Carnegie Mellon University

MELLON COLLEGE OF SCIENCE

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

TITLE: Online Algorithms and Extremal Structures

PRESENTED BY: Joseph Briggs

ACCEPTED BY THE DEPARTMENT OF: Mathematical Sciences

Wesley Pegden	August 2018
MAJOR PROFESSOR	DATE
Thomas Bohman	August 2018
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

Rebecca W. Doerge	August 2018
DEAN	DATE

Online Algorithms and Extremal Structures

Joseph Briggs

Department of Mathematical Sciences

Carnegie-Mellon University

Dissertation Committee

Wesley Pegden, chair
Boris Bukh
Andrzej Dudek
Alan Frieze

A thesis submitted to the Department of Mathematical Sciences
in partial fulfilment for the degree of Doctor of Philosophy
in Algorithms, Combinatorics and Optimization

August 2018

Acknowledgements

I am particularly grateful to my coauthor and advisor Wesley Pegden, who has invested a lot of time and thought into my career and without whom I would not have been able to complete this PhD. I am also very grateful towards Alan Frieze for his patience and guidance with the content of the first two chapters included here. I wish to thank Boris Bukh for his endless support, and whose contribution to my development as a mathematician has been substantial. I am very grateful to Andrzej Dudek for serving on my committee. I furthermore wish to thank Boris Bukh, Tom Bohman, Alan Frieze, Po-Shen Loh and Wesley Pegden for their serious efforts in teaching the combinatorics classes at the graduate level, and in general for fostering a very open and communicative environment ideal for collaborative research. I am grateful to my coauthors Alan Frieze, Michael Krivelevich, Po-Shen Loh, and Benny Sudakov for their joint efforts in maximizing the quality of the content in the first chapter. I am also especially grateful towards fellow graduate students and coauthors Michael Anastos and Chris Cox, whose insight, persistence and determination were crucial for the completion and clarity of the second and third chapters.

It gives me great pleasure to acknowledge the positive mathematical influence on me from endless discussions with past and present students, Michael Anastos, Debsoumya Chakraborti, Daqi Chen, Chris Cox, Mihir Hasabnis, Jennifer Iglesias, Zilin Jiang, Tony Johansson, Misha Lavrov, and Andy Zucker. I am also grateful to fellow graduate students Giovanni Gravina, Will Gunther, Greg Kehne, Brian Kell, Clive Newstead, Kevin Ou, Ilqar Ramazanli, Daniel Rodriguez, Oleksandr Rudenko, Anish Sevekari, Marla Slusky, and Son Van, as well as friends Ilona Ambartsumyam, Tiffany Bao, Dana Bartosova, Quinn Donahoe, Chris Kapulkin, Eldar Khattatov, Youngmin Park, Jay Pina, and Peilun Xi, for all the forms of help and encouragement offered during the last 5 years.

I am very thankful to my family for their support in my further study, and specifically to my sister Katy and my parents for taking the time to visit me from the UK.

Finally, I would like to thank Xiao Chang for her continuous encouragement and belief in my abilities through the times, both good and bad.

Chapter 2 is a version of the paper *Packing Hamilton Cycles Online*, which has been accepted to Combinatorics, Probability, and Computing, and was coauthored with Alan Frieze, Michael Krivelevich, Po-Shen Loh, and Benny Sudakov. Chapter 3 is a version of the paper *Packing Directed and Hamilton Cycles Online*, coauthored with Michael Anastos. Chapter 4 is a version of the paper *Inverting the Turán Problem*, coauthored with Chris Cox, and carried out under the supervision of Wesley Pegden. Chapter 5 is a version of the paper *Extremal Collections of k-Uniform Vectors*, coauthored with Wesley Pegden. Chapters 3-5 have all been submitted.

Abstract

We study several problems in probabilistic and extremal combinatorics.

Probabilistic combinatorics is the area of mathematics studying the behaviour of "most" discrete structures in a given family, as opposed to extremal combinatorics, where one is concerned with the behaviour of "the best" structure with respect to a discrete parameter.

The largest branch of probabilistic combinatorics is the field of random graphs. Given a probability distribution on a collection of graphs, we wish to determine whether a graph-theoretic property is likely to arise or not.

We first study the property of containing q disjoint Hamilton cycles, for which having minimum-degree $2q$ is already known to be sufficient for graphs generated using the random graph process. We establish a dynamic variant, where edges have to be colored online with one of q colors as they appear; with no foresight into the randomness of the edges not yet revealed. Next, we establish a corresponding theorem in the case of directed graphs.

A central theme in extremal combinatorics is the study of the Turán number $\mathrm{ex}(K_n, \mathcal{H})$ of a family of graphs $\mathcal{H}$; the largest number of edges among all $\mathcal{H}$-free subgraphs F of the complete graph K_n. The second problem concerns replacing K_n with a more general host graph G, and asks how many edges G can have as a function of $\mathrm{ex}(G, \mathcal{H})$. We establish many asymptotic and structural results for different $\mathcal{H}$, and show how many previous results arise as special cases of this problem. We also illustrate the robustness of this question in multigraph and non-uniform hypergraph settings.

The third problem concerns matrices M whose columns are distinct and have exactly k nonzero entries, which arise naturally when generating random representable matroids. Given that the rank is some $r \geq 2^{\Omega(k^2)}$, we show the matrix with the most columns has only r or $r+1$ rows, and that this matrix is unique. For finite fields, we also show a corresponding result when the columns have exactly k nonzero entries and the matrix has rank $r \geq \Omega(k^{3/2})$.

Contents

List of Figures

Chapter 1

Introduction

A Hamilton Cycle in a graph G is simply a closed path going through each vertex exactly once, never reusing any edge. They are of fundamental interest in the field of operations research, where they provide efficient ways of moving around every vertex in commonly arising instances of graphs. Thus, many important questions arise of the form: "When does a graph have a Hamilton Cycle?" Certainly, it is necessary for every vertex v to have ≥ 2 edges, where we say the graph has *minimum degree* $\delta(G) \geq 2$.

Since many such networks are both large and dynamic in nature, it is natural to ask structural questions about random graphs. A standard way of constructing a random graph is by repeatedly adding one edge at a time, chosen uniformly at random among the pairs of vertices not currently connected with an edge. After m steps, we obtain a random graph $G_{n,m}$ distributed uniformly among all m-edge graphs with n vertices.

A difficult question dating back to Erdős asked how large an m is needed for $G_{n,m}$ to have a Hamilton cycle with probability $1 - o(1)$ (w.h.p.). A lot of work determining the correct asymptotics of m eventually culminated in the beautiful theorem due to Bollobás [9] and Ajtai, Komlós and Szemerédi [3] stating that as soon as every vertex in $G_{n,m}$ has ≥ 2 edges, there is a Hamilton cycle w.h.p. A classical result due to Dirac tells us that you need $\delta(G) \geq \frac{n}{2}$ in the worst case, but this theorem tells us that most graphs are not "the worst case", in a very strong fashion. In fact, this was strengthened further by Bollobás and Frieze [12] showing that as soon as every vertex has $\geq 2\sigma$ edges, there are actually σ edge-disjoint Hamilton cycles, for *any* fixed choice of σ. That is, we can paint the edges of $G_{n,m}$ in σ different colors so that there is a Hamilton cycle in every color.

The only drawback of this strengthening is that the coloring pays no regard to the dynamic nature of the random graph process. In Chapter 2, we prove an online version, where the edges have to be colored as soon as they appear, without knowledge of which random edges would be appearing next.

In Chapter 3, we obtain an equivalent result for *directed* graphs, where all edges are necessarily one-way streets (but 2 vertices may now have 2 edges connecting them, one in

each direction), using a more complicated algorithm. These results make up some of the strongest evidence yet for the minimum degree being the only obstacle for Hamilton cycles in the random graph setting.

In extremal combinatorics, we study how a parameter is optimized over various families of discrete structures, and either attempt to determine the asymptotic growth of this parameter (according to some fixed notion of size among the families), or when this can be determined precisely, w I also pursue a strong interest in Turán theory. For a fixed graph H (or family thereof), this concerns the maximum number $\mathrm{ex}(n, H)$ of edges in $H-$free graphs on n vertices. Cornerstone theorems here include that of Erdős-Stone, telling us $\mathrm{ex}(n, H)$ is determined asymptotically by the smallest number of colors needed to give adjacent vertices of H distinct colors, when this number is ≥ 3. For bipartite H, however, questions of this type appear very difficult and constitute a very active area of research in their own right (see, for example, [28]).

Approaches in this area typically only look at how large $H-$free graphs that are themselves bipartite can be, since this can only differ from $\mathrm{ex}(n, H)$ by a factor of at most 2. This is effectively a change of host graph from the complete graph K_n to the complete bipartite graph $K_{n/2,n/2}$. In this light, one may similarly ask about the size $\mathrm{ex}(G, H)$ of a maximum $H-$free subgraph of an arbitrary host graph G, and ascribe this parameter to G. Minimizing this over another graph parameter would yield graphs G that are in a sense "best" at forcing copies of the subgraph H, and in particular would have to contain many copies of H. In Section 4, we consider this problem for when the number of edges $|E(G)|$ is fixed. Equivalently, we wish to compute the discrete max-min objective $\mathcal{E}_k(H) := \max\{|E(G)| : \mathrm{ex}(G, H) < k\}$. This is still a general version of problems related to those considered in each of [24], [1], and [5] (where, respectively, $H = \{P_3, K_3\}, P_1 \cup P_2$, and K_n; where P_t is the path with t edges).

It appears, in this setting, that the optimal host graph depends highly upon the choice of H: in these 3 examples, we show the optimal host graphs G are complements of matchings, odd Cayley graphs, and complete graphs respectively. There are results that are not so surprising, such as complete bipartite graphs $K_{n,n}$ being optimal at forcing the collection of even cycles $\{C_4, C_6, \dots\}$. But it is also possible to ask questions of this type in a very general setting, including multigraphs: we can also define $\mathcal{E}_k^*(H) := \max\{|E(G)| : G \textit{ multigraph, } \mathrm{ex}(G, H) < k\}$. In this light, we use the probabilistic method to obtain a surprising result for $\mathcal{O}_2$, the 3-edge graph consisting of a single edge with a loop at each end. Specifically, $\mathcal{E}_k(\mathcal{O}_2) = \frac{3k}{2}$, whereas $\mathcal{E}_k^*(\mathcal{O}_2) \sim \phi k$, where $\phi = 1.618\dots$ is the golden ratio. This illustrates how the corresponding multigraph parameter $\mathcal{E}_k^*$ is not only different from $\mathcal{E}_k$, but also interesting to study in its own right.

Many natural and interesting extremal questions also arise in the context of linear algebra over a finite field $\mathbb{F}_q$. One of the most natural restrictions to place on a collection of vectors is to fix the weight k, i.e. the number of *nonzero* entries. (Indeed, over $\mathbb{F}_2$, such a collection of vectors form the edge-vertex incidence matrix of a k-uniform hypergraph, and when $k = 2$, we obtain precisely the graphic matroid.) Or, in a complementary fashion, one may fix the number k of *zero* entries. The corresponding extremal problems are then to determine the maximum number $\mathrm{ex}(r, k)$ (respectively, $\bar{\mathrm{ex}}(r, k)$) of such columns in a matrix of rank r: for

matrices and matroids, rank is a natural parameter bounding the possible "size".

In Chapter 5, we show that for $r \geq R_k$ sufficiently large with respect to k, $\mathrm{ex}(r,k) = \binom{r}{k}(q-1)^k$ and $\bar{\mathrm{ex}}(r,k) = \binom{r}{k}(q-1)^{r-k}$. In both cases, the bounds are matched by taking all vectors in $\mathbb{F}_q^r$ of weight k (respectively, of weight $r-k$), and in fact these are (eventually) the unique cases attaining equality. However, while the latter result is known once $r \geq \Omega(k^{3/2})$, the proof of the former involves an induction for which a base case can only be indirectly established, and is only known once $r \geq 2^{\Omega(k^2)}$. Nonetheless, this appears to be the greatest progress towards answering a question of Ahlswede, Aydinian, and Khachatrian [2] to classify $\mathrm{ex}(r,k)$ for every $r \geq k$. In particular, it is conjectured that the $2^{\Omega(k^2)}$ above can be replaced by $2k$.

We will see the history and motivation behind each of these problems developed in full in their corresponding chapters' respective introductions.

Chapter 2

Coloring Hamilton Cycles Online

2.1 Introduction

The celebrated random graph process, introduced by Erdős and Rényi [21] in the 1960's, begins with an empty graph on n vertices, and at every step $t = 1, \ldots, \binom{n}{2}$ adds to the current graph a single new edge chosen uniformly at random out of all missing edges. Taking a snapshot of the random graph process after m steps produces the distribution $G_{n,m}$. An equivalent "static" way of defining $G_{n,m}$ would be: choose m edges uniformly at random out of all $\binom{n}{2}$ possible ones. One advantage in studying the random graph process, rather than the static model, is that it allows for a higher resolution analysis of the appearance of monotone graph properties (a graph property is monotone if it is closed under edge addition).

A Hamilton cycle of a graph is a simple cycle that passes through every vertex of the graph, and a graph containing a Hamilton cycle is called Hamiltonian. Hamiltonicity is one of the most fundamental notions in graph theory, and has been intensively studied in various contexts, including random graphs. The earlier results on Hamiltonicity of random graphs were obtained by Pósa [42], and Korshunov [35]. Improving on these results, Bollobás [9], and Komlós and Szemerédi [34] proved that if $m' = \frac{1}{2}n \log n + \frac{1}{2}n \log\log n + \omega n$, then $G_{n,m'}$ is Hamiltonian w.h.p. Here ω is any function of n tending to infinity together with n. One obvious necessary condition for the graph to be Hamiltonian is for the minimum degree to be at least 2, and the above result indicates that the events of being Hamiltonian and of having all degrees at least two are indeed bundled together closely. Bollobás [9], and independently, Ajtai, Komlós, and Szemerédi [3], further strengthened this by proving that w.h.p. the random graph process becomes Hamiltonian when the last vertex of degree one disappears. A more general property $\mathcal{H}_\sigma$ of having σ edge disjoint Hamilton cycles was studied by Bollobás and Frieze [12]. They showed that if $\sigma = O(1)$ then w.h.p. the random graph process satisfies $\mathcal{H}_\sigma$ when the minimum degree becomes 2σ. It took quite a while, but this result was extended to the more difficult case of growing σ in the $G_{n,m}$ context by Knox, Kühn and Osthus [33] and Krivelevich and Samotij [38].

Recently, quite a lot of attention and research effort has been devoted to controlled random graph processes. In processes of this type, an input graph or a graph process is usually generated fully randomly, but then an algorithm has access to this random input and can manipulate it in some well defined way (say, by dropping some of the input edges, or by coloring them), aiming to achieve some preset goal. There is usually the so-called *online* version where the algorithm must decide on its course of action based only on the history of the process so far and without assuming any familiarity with future random edges. For example, in the so-called Achlioptas process the random edges arrive in batches of size k. An online algorithm chooses one of them and puts it into the graph. By doing this one can attempt to accelerate or to delay the appearance of some property. Hamiltonicity in Achlioptas processes was studied in [37]. Another online result on Hamiltonicity was proved in [39]. There, it was shown that one can orient the edges of the random graph process so that w.h.p. the resulting graph has a directed Hamilton cycle exactly at the time when the underlying graph has minimum degree two.

Here we consider a Ramsey-type version of controlled random processes. In this version, the incoming random edge, when it is exposed, is irrevocably colored by an algorithm in one of r colors, for a fixed $r \geq 2$. The goal of the algorithm is to achieve or to maintain a certain monotone graph property in all of the colors. For example, in [8] the authors considered the problem of creating a linear size (so-called *giant*) component in every color.

The above mentioned result of Bollobás and Frieze [12] gives rise to the following natural question. Can one typically construct σ edge disjoint Hamilton cycles in an online fashion by the time the minimum degree becomes 2σ? We answer this question affirmatively in the case $\sigma = O(1)$.

Theorem 2.1.1. *For a fixed integer $\sigma \geq 2$, let $\tau_{2\sigma}$ denote the hitting time for the random graph process $G_i, i = 1, 2, \ldots$ to have minimum degree 2σ. Then w.h.p. we can color the edges of $G_i, i = 1, 2, ..$ online with σ colors so that $G_{\tau_{2\sigma}}$ contains σ Hamilton cycles $C_1, C_2, \ldots, C_\sigma$, where the edges of cycle C_j all have color j.*

2.2 Description of the coloring procedure

We describe our coloring procedure in terms of $q = 2\sigma$ colors we aim to color the edges so that each vertex has degree at least one in each color. Think of colors 1 and $1 + \sigma$ being light red and dark red, say, and then that each vertex is incident with at least two red edges. This may appear cumbersome, but it does make some of the description of the analysis a little easier.

In the broadest terms, we construct two sets of edges E^+ and E^*. Let Γ^*_c be the subgraph of $G_{\tau_{2\sigma}}$ induced by the edges of color c in E^*. We ensure that w.h.p. this has minimum degree at least one for all c. We then show that w.h.p. after merging colors c and $c + \sigma$ for $c \in [\sigma]$ the subgraph $\Gamma^{**}_c = \Gamma^*_c \cup \Gamma^*_{c+\sigma}$ has sufficient expansion properties so that standard arguments using Pósa rotations can be applied. For every color c, the edges of E^*_c are used

to help create a good expander, and produce a backbone for rotations. And the edges in E_c^+ are used to close cycles in this argument.

Notation 2.2.1. "At time t" is taken to mean "when t edges have been revealed".

Notation 2.2.2. Let $N^{(t)}(v)$ denote the set of neighbors of v in G_t and let $d_v^{(t)} = |N^{(t)}(v)|$.

For color $c \in [q]$, write $d_c = d_{c,t}, N_c = N_{c,t}$ for the degrees and neighborhoods of vertices and sets in Γ_c.

Definition 2.2.3. Let $Full$ denote the set of vertices with degree at least $\frac{\epsilon \log n}{1000q}$ in every color at time

$$t_\epsilon := \epsilon n \log n \,,$$

where ϵ is some sufficiently small constant depending only on the constant q. The actual value of ϵ needed will depend on certain estimates below being valid, in particular equation (2.14). A vertex is $Full$ if is lies in $Full$. Similarly, let $Full' \subseteq Full$ denote the set of vertices with degree at least $\frac{\epsilon \log n}{1000q}$ in every color at time $\frac{1}{2}\epsilon n \log n$.

This definition only makes sense if t_ϵ is an integer. Here and below we use the following convention. If we give an expression for an integer quantity that is not clearly an integer, then rounding the expression up or down will give a value that can be used to satisfy all requirements.

2.2.1 Coloring Algorithm COL

We now describe our algorithm for coloring edges as we see them. At any time t, vertex v has a list $C_v^{(t)} := \{c \in [q] : d_c^{(t)}(v) = 0\}$ of colors currently not present among edges incident to v; "the colors that v needs". A vertex is *needy* at time t if $C_v^{(t)} \neq \emptyset$. If the next edge to color contains a needy vertex then we try to reduce the need of this vertex. Otherwise, we make choices to guarantee expansion in E^*, needed to generate many endpoints in the rotation phase, and to provide edges for E^+, which are used to close cycles, if needed.

FOR $t = 1, 2, \ldots, \tau_q$ **DO**
BEGIN

Step 1 Let $e_t = uv$.

Step 2 If $C_v^{(t)} \cup C_u^{(t)} = \emptyset$, $t > t_\epsilon$, and precisely one of $\{u, v\}$ (WLOG u) is $Full$, then give uv the color c that minimises $d_c(v)$ (breaking ties arbitrarily). Add uv to E_c^*.

Step 3 If $C_v^{(t)} \cup C_u^{(t)} = \emptyset$, $t > t_\epsilon$ and *both* $u, v \in Full$, give uv a color c uniformly at random from $[q]$. Then add this edge to E_c^+ or E_c^*, each with probability $1/2$.

Step 4 If $C_v^{(t)} \cup C_u^{(t)} = \emptyset$ but $t \le t_\epsilon$ or *both* $u, v \notin Full$, then color uv with color c chosen uniformly at random from $[q]$. Add uv to E_c^*.

Step 5 Otherwise, color uv with color c chosen uniformly at random from $C_u^{(t)} \cup C_v^{(t)}$. Add uv to E_c^*.

END

Let

$$E^* = \bigcup_{c\in[q]} E_c^* \text{ and } E^+ = \bigcup_{c\in[q]} E_c^+.$$

2.3 Structural properties

Let

$$p = \frac{\log n + (q-1)\log\log n - \omega}{n} \text{ and } m = \binom{n}{2} p$$

where

$$\omega = \omega(n) \to \infty, \omega = o(\log\log n).$$

We will use the following well-known properties relating $G_{n,p}$ and $G_{n,m}$, see for example [26], Chapter 1. Let $\mathcal{P}$ be a graph property. It is monotone increasing if adding an edge preserves it, and is monotone decreasing if deleting an edge preserves it. We have:

$$\mathbb{P}(G_{n,m} \in \mathcal{P}) \le 10m^{1/2}\mathbb{P}(G_{n,p} \in \mathcal{P}). \tag{2.1}$$

$$\mathbb{P}(G_{n,m} \in \mathcal{P}) \le 3\mathbb{P}(G_{n,p} \in \mathcal{P}), \text{ if } \mathcal{P} \text{ is monotone.} \tag{2.2}$$

A vertex $v \in [n]$ is *small* if its degree $d(v)$ in $G_{n,m}$ satisfies $d(v) < \frac{\log n}{100q}$. It is *large* otherwise. The set of small vertices is denoted by $SMALL$ and the set of large vertices is denoted by $LARGE$.

Definition 2.3.1. A subgraph H of $G_{n,m}$ with a subset $S(H) \subset V(H)$ is called a small structure if

$$|E(H)| + |S(H)| - |V(H)| \ge 1.$$

We say that $G_{n,m}$ *contains* H if there is an injective homomorphism $\phi : H \hookrightarrow G_{n,m}$ such that $\phi(S(H)) \subseteq SMALL$. The important examples of H include:

- A single edge between 2 *small* vertices.
- A path of length at most five between two *small* vertices.
- A copy of C_3 or C_4 with at least one *small* vertex.
- Two distinct triangles sharing at least one vertex.

Lemma 1. *For any fixed small structure H of constant size,*

$$\mathbb{P}(G_{n,m} \text{ contains } H) = o(n^{-1/5}).$$

Proof. We will prove that

$$\mathbb{P}(G_{n,p} \text{ contains } H) = o(n^{-3/4}).$$

This along with (2.1) implies the lemma.

Let $h = |V(H)|, f = |E(H)|, s = |S(H)|$ so that $f + s \geq h + 1$. Then:

$$\begin{aligned}
\mathbb{P}\left(G_{n,p} \text{ contains } H\right) &\leq \binom{n}{h} h! p^f \left(\sum_{i=0}^{\frac{\log n}{100q}} \binom{n-h}{i} p^i (1-p)^{n-h-i} \right)^s \\
&\lesssim n^h \left(\frac{\log n}{n}\right)^f \left(\sum_{i=0}^{\frac{\log n}{100q}} \left(\frac{(e+o(1))\log n}{i}\right)^i e^{-\log n - (q-1)\log\log n + \omega + o(1)} \right)^s \\
&\leq n^h \left(\frac{\log n}{n}\right)^f \left(\frac{(300q)^{\frac{\log n}{100q}}}{n(\log n)^{q-1-o(1)}} \right)^s \\
&= o(n^{h-f-s+1/4}) = o(n^{-3/4}).
\end{aligned}$$

(We used the notation $A \lesssim B$ in place of $A \leq (1 + o(1))B$.) In the calculation above, in the first line we placed the vertices of H and decided about the identity of s vertices falling into $SMALL$, then required that all f edges of H are present in $G_{n,p}$, and finally required that for each of the s vertices in $SMALL$, their degree outside the copy of H is at most $\frac{\log n}{100q}$.

□

Lemma 2. *W.h.p., for every $k \in \left[q-1, \frac{\log n}{100q}\right]$, there are less than $\nu_k = \frac{e^{2\omega}(\log n)^{k-q+1}}{(k-1)!}$ vertices of degree k in $G_{n,m}$.*

Remark 2.3.2. ν_k is increasing in k for this range, and for the largest $k = \frac{\log n}{100q}$ we have $\nu_k \lesssim n^{\frac{\log(100eq)}{100q}}$.

Proof. Fix k and then we have

$$\begin{aligned}
&\mathbb{P}(G_{n,p} \text{ has at least } \nu_k \text{ vertices of degree at most } k)\\
&\le \binom{n}{\nu_k}\left(\sum_{\ell=0}^{k}\binom{n-\nu_k}{\ell}p^\ell(1-p)^{n-\nu_k-\ell}\right)^{\nu_k}\\
&= \binom{n}{\nu_k}\left((1+o(1))\binom{n-\nu_k}{k}p^k(1-p)^{n-\nu_k-k}\right)^{\nu_k}\\
&\le \left(\frac{ne}{\nu_k}\times\frac{n^k}{k!}\left(\frac{\log n+(q-1)\log\log n-\omega}{n}\right)^k e^{-\log n-(q-1)\log\log n+\omega+o(1)}\right)^{\nu_k}\\
&\le \left(\frac{e^{\omega+O(1)}}{(\log n)^{q-1}}\frac{(\log n+q\log\log n)^k}{k!\nu_k}\right)^{\nu_k}\\
&= \left(\frac{e^{-\omega+O(1)}}{k}\left(1+\frac{q\log\log n}{\log n}\right)^k\right)^{\nu_k}\\
&\le \left(e^{-\omega+O(1)}\frac{(\log n)^{kq/\log n}}{k}\right)^{\nu_k}.
\end{aligned}$$

The function $f(k)=\frac{(\log n)^{kq/\log n}}{k}$ is log-convex, and so f is maximised at the extreme values of k (specifically $f(q-1)=e^{O(1)}>f\left(\frac{\log n}{100q}\right)=o(1)$). Hence,

$$\mathbb{P}(\exists k: G_{n,p} \text{ has at least } \nu_k \text{ vertices of degree } k)\le\sum_{k=q-1}^{\frac{\log n}{100q}}e^{-\omega\nu_k/2}=o(1).$$

Applying (2.2) we see that

$$\mathbb{P}(\exists k: G_{n,m} \text{ has at least } \nu_k \text{ vertices of degree } k)=o(1),$$

which is stronger than required. □

Lemma 3. *With probability $1-o(n^{-10})$, $G_{n,m}$ has no vertices of degree $\ge 20\log n$.*

Proof. We will prove that w.h.p. $G_{n,p}$ has the stated property. We can then obtain the lemma by applying (2.2).

$$\begin{aligned}
\mathbb{P}(\exists v: d(v)\ge 20\log n) &\le n\binom{n-1}{20\log n}p^{20\log n}\\
&\le n\left(\frac{en}{20\log n}\frac{2\log n}{n}\right)^{20\log n}\\
&\le n\left(\frac{e}{10}\right)^{20\log n}\\
&= o(n^{-10}).
\end{aligned}$$

□

2.4 Analysis of COL

Let $\Gamma = G_m$ and let $d(v)$ denote the degree of $v \in [n]$ in Γ. Let

$$\theta_v = \begin{cases} 0 & d(v) \geq q. \\ 1 & d(v) = q-1. \end{cases}$$

Lemma 4. *Suppose we run COL as described above. Then w.h.p. $|C_v^{(m)}| = \theta_v$ for all $v \in [n]$.*

In words, Lemma 4 guarantees that the algorithm COL typically performs so that at time m, each vertex of degree at least q has all colors present at its incident edges, while each vertex of degree $q-1$ has exactly one color missing. (It is well known that w.h.p. $\delta(G_m) = q-1$, see for example [26], Section **4.2**.)

Proof. Fix v and suppose v has k neighbours in LARGE, via edges $\{f_i = vu_i\}_{i=1}^k$. Then in general $d(v)-1 \leq k \leq d(v)$ as small vertices do not share a path of length two. Also, when v is small, $k = d(v)$. Write $t(e)$ for the time $t \in [1,m]$ at which an edge e appears in the random graph process, i.e. $t(e_i) = i$. Let $t_i = t(f_i)$ and assume that $t_i < t_{i+1}$ for $i > 0$. We omit $i = 1$ in the next consideration since v will always get a color it needs by time t_1. (It may get a color before t_1 through an edge vw where w is not in LARGE.) Every time an $f_i, i \geq 2$, appears while u_i needs no additional colors, v gets a color it needs. So for v to have $|C_v^{(m)}| > \theta_v$ at the end of the process, this must happen at most $q-2-\theta_v$ times, so there is certainly some set

$$S = \{i_1 < i_2 < \cdots < i_s\} \subseteq [2,k] \text{ of } s = k-q+1+\theta_v \text{ indices},$$

whose corresponding edges $\{f_i, i \in S\}$ incident with v satisfy $C_{u_i}^{(t_i)} \neq \emptyset$. Let $\mathbf{T_S}$ denote $\{t_i : i \in S\}$ and $\mathbf{U}$ denote the sequence $u_1, u_2, \ldots, u_k$. In the following we will sum over S and condition on the choices for $\mathbf{T_S}$ and then estimate the probability that $C_{u_i}^{(t_i)} \neq \emptyset$ for $i \in S$. For a fixed S there will be at least $\binom{m-k}{|S|+1}$ equally likely choices for the set $\{t_i, i \in \{1\} \cup S\}$. (We do not condition on t_1. The factor $t_{i_1} - 1$ in (2.3) below will allow for the number of choices for t_1.) Let $\mathcal{L}$ denote the occurrence of the bound of $20\log n$ on the degree of v and its neighbors (see Lemma 3), and note that $\mathbb{P}(\mathcal{L}) = 1 - o(n^{-10})$.

Taking a union bound over S, and letting

$$A_i := \left\{ C_{u_i}^{(t_i)} \neq \emptyset \right\},$$

we have

$$\mathbb{P}(|C_v^{(m)}| > \theta_v \mid \mathcal{L}, \mathbf{U}) \leq \sum_{\substack{S \subset [2,k] \\ |S|=s}} \sum_{t_i : i \in \{1\} \cup S} \frac{1}{\binom{m-k}{k-q+2+\theta_v}} \mathbb{P}\left(\bigwedge_{i \in S} A_i \middle| \mathbf{T_S}, \mathbf{U}, \mathcal{L} \right)$$

$$\approx \sum_{\substack{S\subset[2,k]\\|S|=s}} \sum_{t_i:i\in S} \frac{t_{i_1}-1}{\binom{m}{k-q+2+\theta_v}} \mathbb{P}\Big(\bigwedge_{i\in S} A_i \Big| \mathbf{T_S}, \mathbf{U}, \mathcal{L} \Big), \tag{2.3}$$

since there are $t_{i_1}-1$ choices for t_1 and $k^2=o(m)$, implying $\binom{m-k}{k-q+2+\theta_v} \approx \binom{m}{k-q+2+\theta_v}$, given $\mathcal{L}$. Next let

$$\begin{aligned}
&Y_i = \{\text{edges of } u_i \text{ that appeared before } t_i \text{ excluding edges contained in } N^{(m)}(v)\},\\
&d_r = d(u_r) \text{ and } Z_r := |Y_r| \text{ for } r = 1,2,\dots,s,\\
&\mathbf{D_S} = \{d_i : i \in S\}.
\end{aligned}$$

Now fix $\mathbf{U}$ and S and $\mathbf{T_S}$ and $\mathbf{D_S}$.

Remark 2.4.1. Going back to Algorithm COL, we observe that Step 5 implies that if $C_v^{(t)} \neq \emptyset$ then uv is colored with a color in $C_v^{(t)}$ with probability at least $\frac{1}{q}$. This holds regardless of the previous history of the algorithm and also holds conditional on $\mathbf{T_S}, \mathbf{U}, \mathbf{D_S}$. Indeed, the random bits used in Step 5 are independent of the history and are distinct from those used to generate the random graphs. The latter explains why we can condition on the future by fixing $\mathbf{T_S}, \mathbf{U}, \mathbf{D_S}$. We condition on $\mathcal{L}$ in order to control s as $O(\log n)$.

Then,

$$\begin{aligned}
&\mathbb{P}\left(A_{i_1} \wedge \cdots \wedge A_{i_s} \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}\right)\\
&= \sum_{z_s} \underbrace{\mathbb{P}(A_{i_s} \mid A_{i_1}, \dots, A_{i_{s-1}}, Z_s = z_s, \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})}_{\leq \mathbb{P}(\mathrm{Bin}(z_s, q^{-1}) \leq q-1) \text{ by Remark 3.2.5}} \mathbb{P}(A_{i_1}, \dots, A_{i_{s-1}}, Z_s = z_s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})\\
&\leq \sum_{z_s} g(z_s) \sum_{z_{s-1}} \mathbb{P}(A_{i_{s-1}} \mid A_{i_1}, \dots, A_{i_{s-2}}, Z_{s-1} = z_{s-1}, Z_s = z_s, \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})\\
&\qquad \times \mathbb{P}(A_{i_1}, \dots, A_{i_{s-2}}, Z_{s-1} = z_{s-1}, Z_s = z_s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})\\
&\leq \sum_{z_s, z_{s-1}} g(z_s) g(z_{s-1}) \mathbb{P}(A_{i_1}, \dots, A_{i_{s-2}}, Z_{s-1} = z_{s-1}, Z_s = z_s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})\\
&\leq \sum_{z_s, \dots, z_1} g(z_s) \cdots g(z_2) \mathbb{P}(Z_r = z_r, r = 1, \dots, s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}) \text{ (by induction)}
\end{aligned} \tag{2.4}$$

Here $g(z) := \mathbb{P}(\mathrm{Bin}(z, q^{-1}) \leq q-1)$ for any $z \geq 0$.

Claim 2.4.2.

$$\mathbb{P}(Z_r = z_r, r = 1, 2, \dots, s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}) \leq \left(1 + \tilde{O}(n^{-1})\right) \prod_{r=1}^{s} \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}},$$

where $\tilde{O}$ hides polylog factors.

Proof Fix $\frac{\log n}{100q} \le d_1, d_2, \dots, d_s = O(\log n)$ and $t_1, t_2, \dots, t_s$. Then, for every $1 \le r \le s$,

$$\mathbb{P}(Z_r = z_r \mid Z_{r-1} = z_{r-1}, \dots, \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}) \le (1 + o(n^{-10})) \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r - z_r}}{\binom{m-d_2-\cdots-d_{r-1}-s}{d_r}} \tag{2.5}$$

$$\le \left(1 + \tilde{O}(n^{-1})\right) \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}}.$$

Explanation for (2.5)**:** The the first binomial coefficient in the numerator in (2.5) bounds the number of choices for the z_r positions in the sequence where an edge contributing Y_r occurs. This holds regardless of $z_1, z_2, \dots, z_{r-1}$. The second binomial coefficient bounds the number of choices for the $d_r - z_r$ positions in the sequence where we choose an edge incident with u_r after time t_r. Conversely, the denominator in (2.5) is a lower bound on the number of choices for the d_r positions where we choose an edge incident with u_r, given $d_1, d_2, \dots, d_{r-1}$. We subtract the extra s to (over)count for edges from v to $u_{r+1}, \dots, u_s$. The factor $(1 + o(n^{-10}))$ accounts for the conditioning on $\mathcal{L}$.

Expanding $\mathbb{P}(Z_r = z_r, r = 1, \dots, s \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L})$ as a product of $s = O(\log n)$ of these terms completes the proof of Claim 2.4.2. □

Going back to (2.4) we see that given $d_1, d_2, \dots, d_s$,

$$\begin{aligned}
&\mathbb{P}\left(A_{i_1} \wedge \cdots \wedge A_{i_s} \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}\right) \\
&\lesssim \prod_{r=1}^{s} \sum_{z_r=0}^{d_r} \left(\mathbb{P}(Bin(z_r, q^{-1}) \le q-1) \times \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} \right) \\
&\le \prod_{r=1}^{s} \sum_{z_r=0}^{d_r} \left(C_1 \binom{z_r}{\min\{z_r, q-1\}} \frac{1}{q^{q-1}} \left(1 - \frac{1}{q}\right)^{z_r} \times \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} \right) \qquad (2.6) \\
&\le \prod_{r=1}^{s} \sum_{z_r=0}^{d_r} \left(C_1 \max\left\{1, z_r^{q-1}\right\} e^{-z_r/q} \times \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} \right). \qquad (2.7)
\end{aligned}$$

Here, $C_1 = C_1(q)$ depends only on q. We will use constants $C_2, C_3, \dots$ in a similar fashion without further comment.

Justification for (2.6)**:** If $z_r \le q-1$ then $\mathbb{P}(Bin(z_r, q^{-1}) \le q-1) = 1$ and $C_1 = eq^q$ will suffice.

If $q \le z_r \le 10q$ we use

$$\mathbb{P}(Bin(z_r, q^{-1}) \le q-1) \le 1 \text{ and } \binom{z_r}{q-1} \frac{1}{q^{q-1}} \left(1 - \frac{1}{q}\right)^{z_r} \ge \frac{1}{q^{q-1}} \left(1 - \frac{1}{q}\right)^{10q}$$

and $C_1 = e^{20}q^q$ will suffice in this case.

If $z_r > 10q$ then putting $a_i := \mathbb{P}(Bin(z_r, q^{-1}) = i) = \binom{z_r}{i}\frac{1}{q^i}\left(1-\frac{1}{q}\right)^{z_r-i}$ for $i \le q-1$ we see that

$$\frac{a_i}{a_{i-1}} = \frac{z_r - i + 1}{i} \cdot \frac{1}{q-1} \ge \frac{z_r - q}{q^2} > \frac{z_r}{2q^2} \ge \frac{5}{q}.$$

So here

$$\mathbb{P}(\mathrm{Bin}(z_r, q^{-1}) \le q-1) = \sum_{i=0}^{q-1} a_i \le a_{q-1}\left(1 + \frac{2q^2}{z_r} + \cdots + \left(\frac{2q^2}{z_r}\right)^{q-2}\right) \le$$
$$\left(1-\frac{1}{q}\right)^{1-q}\left(\binom{z_r}{q-1}\frac{1}{q^{q-1}}\left(1-\frac{1}{q}\right)^{z_r}\right)\frac{\left(\frac{q}{5}\right)^{q-1}-1}{\frac{q}{5}-1},$$

and thus $C_1 = (5q)^q$ suffices.

This completes the verification of (2.6).

Now, writing $(t)_z$ for the *falling factorial* $t!/(t-z)! = t(t-1)(t-2)\dots(t-z+1)$,

$$\begin{aligned}\frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} &= \binom{d_r}{z_r}\frac{(t_r)_{z_r}(m-t_r)_{d_r-z_r}}{(m)_{d_r}} \\ &= \binom{d_r}{z_r}\prod_{i=0}^{z_r-1}\frac{t_r-i}{m-(d_r-z_r)-i}\cdot\prod_{i=0}^{d_r-z_r-1}\frac{m-t_r-i}{m-i} \\ &\le \left(1+O\left(\frac{d_r^2}{m}\right)\right)\binom{d_r}{z_r}\left(\frac{t_r}{m}\right)^{z_r}\left(1-\frac{t_r}{m}\right)^{d_r-z_r}. \qquad (2.8)\end{aligned}$$

Observe next that if $z_r \ge q^2$ then

$$(z_r)_{q-1} = z_r^{q-1}\prod_{i=0}^{q-1}\left(1-\frac{i}{z_r}\right) \ge z_r^{q-1}\left(1-\frac{q^2}{2z_r}\right) \ge \frac{z_r^{q-1}}{2}. \qquad (2.9)$$

It follows from (2.8) and (2.9) that

$$\begin{aligned}&\sum_{z_r=q^2}^{d_r} C_1 z_r^{q-1}e^{-z_r/q} \times \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} \\ &\le 2C_1\sum_{z_r=q-1}^{d_r}(z_r)_{q-1}\binom{d_r}{z_r}\left(\frac{t_re^{-1/q}}{m}\right)^{z_r}\left(1-\frac{t_r}{m}\right)^{d_r-z_r} \\ &\le 2C_1(d_r)_{q-1}\left(\frac{t_r}{m}\right)^{q-1}\sum_{z_r=q-1}^{d_r}\binom{d_r-q+1}{z_r-q+1}\left(\frac{t_re^{-1/q}}{m}\right)^{z_r-q+1}\left(1-\frac{t_r}{m}\right)^{d_r-z_r} \\ &\le 2C_1\left(\frac{d_rt_r}{m}\right)^{q-1}\left(1-\frac{t_r}{m}\left(1-e^{-1/q}\right)\right)^{d_r-q+1}\end{aligned}$$

$$\leq 2C_1 \left(\frac{d_r t_r}{m}\right)^{q-1} \exp\left\{-\frac{(d_r-q+1)t_r}{m}\left(1-e^{-1/q}\right)\right\}.$$

Furthermore, not forgetting

$$\begin{aligned}\sum_{z_r=0}^{q^2-1} C_1 \max\left\{1, z_r^{q-1}\right\} e^{-z_r/q} \times \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} &\leq C_2 \sum_{z_r=0}^{q^2-1} \frac{\binom{t_r}{z_r}\binom{m-t_r}{d_r-z_r}}{\binom{m}{d_r}} \\ &\leq C_3 \sum_{z_r=0}^{q^2-1} t_r^{z_r} \cdot \frac{(m-t_r)^{d_r-z_r}}{(d_r-z_r)!} \cdot \frac{d_r!}{m^{d_r}} \\ &\leq C_3 \sum_{z_r=0}^{q^2-1} \left(\frac{d_r t_r}{m}\right)^{z_r} e^{-(d_r-z_r)t_r/m} \\ &\leq C_4 \psi\left(\frac{d_r t_r}{m}\right),\end{aligned}$$

where $\psi(x) = e^{-x}\sum_{z=0}^{q^2-1} x^z$. (Now $z_r \leq q^2$ and so the factor $e^{z_r t_r/m} \leq e^{q^2}$ can be absorbed into C_4.) Going back to (2.7) we have

$$\mathbb{P}\left(A_{i_1} \wedge \cdots \wedge A_{i_s} \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}\right) \leq C_5^s \prod_{r=1}^{s} \left(\left(\frac{d_r t_r}{m}\right)^{q-1} \exp\left\{-\frac{d_r t_r}{m}\left(1-e^{-1/q}\right)\right\} + \psi\left(\frac{d_r t_r}{m}\right)\right). \quad (2.10)$$

It follows from (2.3) and (2.10) that,

$$\begin{aligned}p_v &:= \mathbb{P}(|C_v^{(m)}| > \theta_v \mid \mathbf{T_S}, \mathbf{U}, \mathbf{D_S}, \mathcal{L}) \\ &\leq \sum_{\substack{S\subset[2,k]\\|S|=s}} \sum_{t_i: i\in S} \frac{t_{i_1} C_5^s}{\binom{m}{s+1}} \prod_{r=1}^{s} \left(\left(\frac{d_r t_r}{m}\right)^{q-1} \exp\left\{-\frac{d_r t_r}{m}(1-e^{-1/q})\right\} + \psi\left(\frac{d_r t_r}{m}\right)\right).\end{aligned}$$

Replacing a sum of products by a product of sums and dividing by $s!$ to account for repetitions, we get

$$\begin{aligned}p_v \leq \sum_{\substack{S\subset[2,k]\\|S|=s}} \frac{C_5^s}{\binom{m}{s+1}s!} \prod_{r=2}^{s} &\left(\sum_{t=1}^{m} \left(\left(\frac{d_r t}{m}\right)^{q-1} \exp\left\{-\frac{d_r t}{m}(1-e^{-1/q})\right\} + \psi\left(\frac{d_r t}{m}\right)\right)\right) \\ &\times \left(\sum_{t=1}^{m} \left(t\left(\frac{d_1 t}{m}\right)^{q-1} \exp\left\{-\frac{d_1 t}{m}(1-e^{-1/q})\right\} + t\psi\left(\frac{d_1 t}{m}\right)\right)\right).\end{aligned}$$

We now replace the sums by integrals. This is valid seeing as the summands have a bounded number of extrema, and we replace C_5 by C_6 to absorb any small error factors.

$$p_v \leq \sum_{\substack{S\subset[2,k]\\|S|=s}} \frac{C_6^s}{\binom{m}{s+1}s!} \prod_{r=2}^{s} \left(\int_{t=0}^{\infty} \left[\left(\frac{d_r t}{m}\right)^{q-1} \exp\left\{-\frac{d_r t}{m}(1-e^{-1/q})\right\} + \psi\left(\frac{d_r t}{m}\right)\right] dt\right)$$

$$\times\left(\int_{t=0}^{\infty}\left(t\left(\frac{d_1t}{m}\right)^{q-1}\exp\left\{-\frac{d_1t}{m}(1-e^{-1/q})\right\}+\psi\left(\frac{d_1t}{m}\right)\right)dt\right)$$

$$=\sum_{\substack{S\subset[2,k]\\|S|=s}}\frac{C_6^s}{\binom{m}{s+1}s!}\prod_{r=2}^{s}\left(\frac{m}{d_r}\int_{x=0}^{\infty}(x^{q-1}e^{-(1-e^{-1/q})x}+\psi(x))dx\right)$$

$$\times\frac{m^2}{d_1^2}\left(\int_{x=0}^{\infty}\left(x^q\exp\left\{-(1-e^{-1/q})x\right\}+x\psi(x)\right)dx\right)$$

$$\le\sum_{\substack{S\subset[2,k]\\|S|=s}}\frac{C_6^s}{\binom{m}{s+1}s!}\cdot\left(\frac{C_7m}{\min_r\{d_r\}}\right)^{s+1}$$

$$\le\frac{C_8^k}{(\log n)^{k-q+2+\theta_v}}.$$

Applying Lemmas 14 and 3 and removing the conditioning on $\mathbf{T_S},\mathbf{U},\mathbf{D_S},\mathcal{L}$ we see that with $k_0=\frac{\log n}{100q}$,

$$\begin{aligned}
&\mathbb{P}(\exists v:|C_v^{(m)}|>\theta_v)\\
&\le\mathbb{P}(\neg\mathcal{L})+\sum_{k=q-1}^{k_0}\frac{e^{2\omega}(\log n)^{k-q+1}}{(k-1)!}\times\frac{C_8^k}{(\log n)^{k-q+2+\theta_v}}+n\sum_{k=k_0}^{20\log n}\frac{C_8^k}{(\log n)^{k-q+2}}\\
&\le o(1)+\frac{e^{2\omega}}{\log n}\sum_{k\ge q-1}\frac{C_8^k}{(k-1)!}+n\sum_{k=k_0}^{20\log n}\frac{C_8^k}{(\log n)^{k/2}}\\
&\le o(1)+\frac{C_9e^{2\omega+C_9}}{\log n}\\
&=o(1).
\end{aligned}$$

□

We show next that at time m, w.h.p. sets of size up to $\Omega(n)$ have large neighbourhoods in every color.

We first prove that typically "large-degree vertices have large degree in every color": let $d_c^*(v)$ denote the number of edges incident with v that COL colors c, except for those edges that are colored in Step 3.

Theorem 1. *There exists $\epsilon=\epsilon(q)>0$ such that w.h.p. on completion of COL every $v\in LARGE$ has $d_c^*(v)\ge\frac{\epsilon\log n}{1000q}$ for all $c\in[q]$.*

Suppose we define a vertex to be $small_c$ if it has $d_c(v)\le\frac{\epsilon\log n}{1000q}$. Theorem 1 says w.h.p. the set of $small_c$ vertices $SMALL_c\subset SMALL$ so that by Lemma 11, G does not contain any $small_c$ structures of constant size. Here a $small_c$ structure is a small structure made up of $small_c$ vertices.

The proof of Theorem 1 will follow from Lemmas 5, 6 and 7 below.

Lemma 5. *There exists $\delta = \delta(q) > 0$ such that the following holds w.h.p.: Let $Full', Full$ be as in Definition 2.2.3. Then $|Full'| \geq n - \frac{203qn}{\epsilon \log n}$, and $|Full| \geq n - n^{1-\delta}$.*

Proof. We first note that for $v \in [n]$, that if $t_{\mathbf{e}} = \epsilon n \log n$ then

$$\mathbb{P}\left(d^{(t_\epsilon/2)}(v) < \lambda_0 := \frac{\epsilon \log n}{100}\right) \leq 3n^{-\epsilon/2} < n^{-\epsilon/3}. \tag{2.11}$$

Indeed, with $p_1 = \frac{t_\epsilon/2}{\binom{n}{2}}$ we see that, in the random graph model G_{n,p_1}:

$$\mathbb{P}\left(d(v) < \lambda_0\right) = \sum_{i=0}^{\lambda_0 - 1} \binom{n}{i} p_1^i (1-p_1)^{n-i} \leq 2\binom{n}{\lambda_0} p_1^{\lambda_0}(1-p_1)^{n-\lambda_0} \leq 2\left(\frac{nep_1}{\lambda_0}\right)^{\lambda_0} n^{o(1)-\epsilon} \leq n^{-\epsilon/2}. \tag{2.12}$$

The first inequality follows from the fact that the ratio of succesive summands in the sum is at least $(n - \lambda_0)p_1/\lambda_0 > 50$.

Equation (2.11) now follows from (2.2) (with p replaced by p_1) and (2.12).

Thus the Markov inequality shows that with probability at least $1 - n^{-\epsilon/3}$, at least $n - n^{1-\epsilon/6}$ of the vertices v have $d^{(t_\epsilon/2)}(v) \geq \frac{\epsilon \log n}{100}$. Now note that at most qn of the first $t_\epsilon/2$ edges were restricted in color by being incident to at least one needy vertex. This is because each time a needy vertex gets an edge incident to it, the total number of needed colors decreases by at least one. Therefore at most $\frac{200qn}{\epsilon \log n}$ of these vertices v have fewer than $\frac{\epsilon \log n}{200}$ of their $\geq \frac{\epsilon \log n}{100}$ initial edges colored completely at random, as in Step 4 of COL. Hence, there are at least $n - \frac{201qn}{\epsilon \log n}$ vertices v which have $d^{(t_\epsilon/2)}(v) \geq \frac{\epsilon \log n}{100}$ and $\frac{\epsilon \log n}{200}$ edges of fully random color. For such a v, and any color c,

$$\mathbb{P}\left(d_c^{(t_\epsilon/2)}(v) < \frac{\epsilon \log n}{1000q}\right) \leq \mathbb{P}\left(Bin\left(\frac{\epsilon \log n}{200}, \frac{1}{q}\right) \leq \frac{\epsilon \log n}{1000q}\right) \leq \exp\left\{\frac{-16 \cdot \epsilon \log n}{50 \cdot 200q}\right\} \leq \frac{1}{n^{\frac{-\epsilon}{1000q}}}. \tag{2.13}$$

So $\mathbb{P}(v \notin Full') \leq qn^{-\epsilon/1000q}$, and the Markov inequality shows that w.h.p.

$$|Full'| \geq n - \frac{201qn}{\epsilon \log n} - n^{1-\epsilon/2000q} \geq n - \frac{202qn}{\epsilon \log n}.$$

Now, for $v \notin Full'$, let $S(v) := Full' \setminus N^{(t_\epsilon/2)}(v)$. Since $d^{(t_\epsilon/2)}(v) \leq d(v) \leq 20 \log n$, we have $|S(v)| \geq n - \frac{203qn}{\epsilon \log n}$. Furthermore, every $w \in S(v)$ is no longer needy, and so among the next $t_\epsilon/2$ edges, at most q of the edges between v and $S(v)$ have their choice of color restricted by v, and the rest are colored randomly as in Step 4 of COL. Now $\mathbb{P}(|e(v, S(v))| < \frac{\epsilon \log n}{100}) = O(n^{-\epsilon/3})$ by a similar calculation to (2.12). Conditioning on $|e(v, S(v))| \geq \frac{\epsilon \log n}{100}$ we have $\mathbb{P}(v \notin Full) \leq qn^{-\epsilon/1000q}$ by a similar calculation to (2.13) and so $|Full| \geq n - n^{1-\epsilon/2000q}$ with probability $\geq 1 - O(n^{-\epsilon/2000q})$ by the Markov inequality. □

We are working towards showing that vertices with low degree in some color must have have a low overall degree. The point is that all $Full$ vertices no longer need additional colors

later than $t_\epsilon = \epsilon n \log n$, so any new edge connecting $Full$ to $V \setminus Full$ after time t_ϵ has its color determined by the vertex not in $Full$, as in Step 2 of COL. Indeed, suppose a vertex $v \notin Full$ has at least $\frac{\epsilon \log n}{400}$ edges to $Full$ after time t_ϵ. Then v gets at least $\frac{\epsilon \log n}{400q} > \frac{\epsilon \log n}{1000q}$ edges of every color incident with it.

Lemma 6. *W.h.p. there are no vertices $v \notin Full$ with at least $\frac{\epsilon \log n}{200}$ edges after time t_ϵ i.e., $d^{(m)}(v) - d^{(t_\epsilon)}(v) \geq \frac{\epsilon \log n}{200}$ but with at most $\frac{\epsilon \log n}{400}$ of these edges to $Full$.*

Proof. Take any vertex $v \notin Full$ and consider the first $\frac{\epsilon \log n}{200}$ edges incident to v after time t_ϵ. We must estimate the probability that at least half of these edges are to vertices not in $Full$. We bound this by

$$\binom{\epsilon \log n/200}{\epsilon \log n/400} \left(\frac{n^{1-\delta}}{n - 20 \log n}\right)^{\epsilon \log n/400} = o(1/n).$$

We subtract off a bound of $20 \log n$ on the number of edges from v to $Full$ in E_{t_ϵ}. Note that we do not need to multiply by the number of choices for $Full$, as $Full$ is defined by the first t_ϵ edges. There at most n choices for v and so the lemma follows. □

Lemma 7. *There are no large vertices v with $d^{(m)}(v) - d^{(t_\epsilon)}(v) < \frac{\epsilon \log n}{200}$.*

Proof. Any v satisfying these conditions must have $d^{(t_\epsilon)}(v) \geq \frac{\log n}{200q}$, if $\epsilon \leq 1/q$ say. However, with $p_2 = \frac{t_\epsilon}{\binom{n}{2}}$ we have that in the random graph model G_{n,p_2},

$$\mathbb{P}\left(d(v) \geq \frac{\log n}{200q}\right) \leq \binom{n}{\log n/200q} p_2^{\log n/200q} \leq (400qe\epsilon)^{\log n/200q} = o(n^{-2}), \tag{2.14}$$

for t_ϵ sufficiently small.

The result follows by taking a union bound over choices of v and using (2.2) (again noting $\binom{n}{2} p_2 = t_\epsilon \to \infty$). □

Proof of Theorem 1: It follows from Lemmas 6, 7 that every large vertex has at least $\frac{\epsilon \log n}{400}$ edges to $Full$ that occur after time t_ϵ. These edges will provide all needed edges of all colors. □

It is known that w.h.p. $m \leq \tau_q \leq m' = m + 2\omega n$, see Erdős and Rényi [**?**]. We have shown that at time m all vertices, other than vertices of degree $q-1$, are incident with edges of all colors. Furthermore, vertices of degree $q-1$ are only missing one color. As we add the at most $2\omega n$ edges needed to reach τ_q we find (see Claim 2.4.3 below) that w.h.p. the edges we add incident to a vertex v of degree $q-1$ have their other end in LARGE. As such COL will now give vertex v its missing color.

Claim 2.4.3. *W.h.p. an edge of $E_{m'} \setminus E_m$ that meets a vertex of degree $q-1$ in G_m has its other end in LARGE.*

Proof. It follows from Lemma 14 that at time m and later there are w.h.p. at most $e^{2\omega}$ vertices of degree $q-1$ and at most

$$M = \sum_{k=q-1}^{\log n/100q} \frac{e^{2\omega}(\log n)^{k-q+1}}{(k-1)!} \leq 2e^{2\omega}\left(\frac{e\log n}{\log n/100q}\right)^{-q+1+\log n/100q} \leq n^{1/3}$$

vertices in SMALL. (The first inequality follows from the fact that the summands grow by a factor of at least $100q$.)

Thus the probability that there is an edge contradicting the claim is at most

$$2\omega n \times \frac{e^{2\omega}\times n^{1/3}}{\binom{n}{2}-m'} = o(1).$$

□

We remind the reader that $q = 2\sigma$ where we only use σ colors. We apply the above analysis by identifying colors *mod* σ. We therefore have the following:

Corollary 2.4.4. *W.h.p. the algorithm COL applied to $G_{\tau_{2\sigma}}$ yields a coloring for which $d_c^*(v) \geq 2$ for all $v \in [n]$.*

Proof. We can see from the above that w.h.p. at time $\tau_{2\sigma}$ we have that $d_c(v) \geq 2$ for all $c \in [\sigma], v \in [n]$. Furthermore, by construction, for each $c \in [q], v \in [n]$ the first edge incident with v that gets color c will be in E_c^*. (The only time we place an edge in E_c^+ is when it joins two full vertices.) □

From now on we think in terms of σ colors.

2.4.1 Expansion

For a set $S \subseteq [n]$ we let

$$N_c^*(S) = \{v \notin S : \exists u \in S \text{ s.t. } uv \in E_c^*\} \subset N_c(S).$$

Let

$$\alpha = \frac{1}{10^6 q}.$$

Lemma 8. *Then w.h.p. $|N_c^*(S)| \geq 19|S|$ for all $S \subset LARGE$, $|S| \leq \alpha n$.*

Claim 2.4.5. *At time m, for every $R \subset V(G)$ with $|R| \leq \frac{n}{(\log n)^3}$, there are w.h.p. at most $2|R|$ edges within R of color c for every $c \in [q]$.*

Proof of Claim: We will show that w.h.p. every such R does not have this many edges irrespective of color. Note that the desired property is monotone decreasing, so it suffices to use (2.2) and show this occurs w.h.p. in $G_{n,p}$:

$$\begin{aligned}
&\mathbb{P}\left(\exists |R| \le \frac{n}{(\log n)^3} : |E(G[R])| > 2|R|\right) \\
&\le \sum_{r=4}^{n/(\log n)^3} \binom{n}{r}\binom{\binom{r}{2}}{2r} p^{2r} \\
&\le \sum_{r=4}^{n/(\log n)^3} \left(\frac{ne}{r}\left(\frac{re^{1+o(1)}\log n}{4n}\right)^2\right)^r \\
&\le \sum_{r=4}^{n/(\log n)^3} \left(\frac{r}{n}\cdot\frac{e^{3+o(1)}(\log n)^2}{16}\right)^r = o(n^{-3}).
\end{aligned}$$

□

Proof of Lemma 8:
Case 1: $|S| \le \frac{n}{(\log n)^4}$.
We may assume that $S \cup N_c^*(S)$ is small enough for Claim 2.4.5 to apply (otherwise $|N_c^*(S)| \ge \frac{n}{(\log n)^3} - \frac{n}{(\log n)^4}$ so that S actually has logarithmic expansion in color c). Then, using e_c to denote the number of edges in color c, and using Theorem 1,

$$\frac{\epsilon \log n}{1000q}|S| \le \sum_{v\in S} d_c^*(v) = 2e_c(S) + e_c(S, N_c^*(S)) \le 4|S| + 2|N_c^*(S)\cup S|.$$

Hence,

$$|N_c^*(S)| \ge \frac{\epsilon \log n}{2001q}|S| \ge 19|S|,$$

which verifies the truth of the lemma for this case.

Case 2: $\frac{n}{(\log n)^4} \le |S| \le \frac{n}{50\log n}$.
Let

$$m_+ := \frac{n\log n}{8q}.$$

Let E_c^+, E_c^* denote the edges of E^+, E^* respectively, which are colored c. We begin by proving

Claim 2.4.6. *$|E_c^+|, |E_c^*| \ge m_+$ w.h.p.*

Proof. Once $Full$ has been formed, it follows from Lemma 5, that at most $(n^{1-\delta}(n-n^{1-\delta}))+\binom{n^{1-\delta}}{2} < 2n^{-\delta}\binom{n}{2}$ spaces remain in $E(Full, V\setminus Full)$ or $E(V\setminus Full)$. For each of the $m - t_\epsilon \sim (\frac{1}{2}-\epsilon)n\log n$ edges appearing thereafter, since $\lesssim n\log n < n^{-\delta}\binom{n}{2}$ edges have been placed already, each has a probability $\ge 1-4n^{-\delta}$ of having both ends in $Full$, independently of what has happened previously. Applying the Chernoff bounds (see for example [26],

Chapter 21.4) we see that the probability that fewer than $\frac{1}{3}n\log n$ of these $(\frac{1}{2}-\epsilon)n\log n$ edges were between vertices in $Full$ is at most $e^{-\Omega(n\log n)}$. We remind the reader that every edge with both endpoints in $Full$ is randomly colored and placed in E^+ or E^* in Step 3 of COL.

So, we may assume there are at least $\frac{1}{3q}n\log n$ of these edges in $E^+\cup E^*$ of color c in expectation and then the Chernoff bounds imply that there are at least $\frac{1}{8q}n\log n = m^+$ w.h.p. in both E^+ and E^*. □

Suppose there exists S as above with $|N_c^*(S)| < \frac{\log n}{1000q}|S|$. For $F := S\cap Full$, note that $|F|\geq |S|-n^{1-\delta} = |S|(1-o(1))$. Therefore $|N_c^*(F)\cap Full| < \frac{\log n}{1000q}|S| \leq \frac{\log n}{999q}|F|$. We will show that w.h.p. there are no such $F\subseteq Full$.

We consider the graphs $H_1 = G_{|Full|,m_+}\setminus E_{t_\epsilon}$ and the corresponding independent model $H_2 = G_{|Full|,p_+}\setminus E_{t_\epsilon}$ where $p_+\sim\frac{\log n}{4qn}$. We will show that w.h.p. H_2 contains no set F of the postulated size and small neighborhood. Together with (2.2) (and $\binom{|Full|}{2}p_+\to\infty$) this implies that w.h.p. H_1 has no such set either. Note that by Lemma 3, we see that w.h.p. at most $20|F|\log n$ edges of E_{t_ϵ} are incident with F. This calculation is relevant because $(E^*\setminus E_{t^*})$'s only dependence on E_{t_ϵ} is that it is disjoint from it.

Hence, in H_2,

$$\begin{aligned}\mathbb{P}(\exists F) &\leq \sum_{f=(n-o(n))/(\log n)^4}^{n/50\log n}\sum_{k=0}^{\frac{\log n}{999q}f}\binom{|Full|}{f}\binom{|Full|}{k}f^k p_+^k(1-p_+)^{(|Full|-k)f-20f\log n}\\ &\leq \sum_{f=(n-o(n))/(\log n)^4}^{n/50\log n}\sum_{k=0}^{\frac{\log n}{999q}f}\underbrace{\left(\frac{ne}{f}\right)^f\left(\frac{nf}{k}\cdot\frac{\log n}{qn}\right)^k e^{-nfp_+(1-o(1))}}_{u_{f,k}}.\end{aligned}$$

Here, the ratio

$$\frac{u_{f,k+1}}{u_{f,k}} = \frac{f\log n}{q(k+1)}\left(\frac{k}{k+1}\right)^k \geq 999/e.$$

Therefore,

$$\mathbb{P}(\exists F)\leq 2\sum_{f\sim n/(\log n)^4}^{n/50\log n}\left(\frac{ne}{f}\cdot(999)^{\frac{\log n}{999q}}n^{-1/5q}\right)^f \leq 2n\left(3(\log n)^4 n^{-1/10q}\right)^{\frac{(1-o(1))n}{(\log n)^4}} = o(1).$$

Case 3: $\frac{n}{50\log n}\leq |S|\leq \frac{n}{10^6 q}$.
Choose any $S_1\subset S$ of size $\frac{n}{50\log n}$, then

$$|N_c^*(S)|\geq |N_c^*(S_1)|-|S| \geq \frac{\log n}{1000q}\cdot\frac{n}{50\log n}-\frac{n}{10^6 q} = 19\alpha n\geq 19|S|.$$

□

The following corollary applies to the subgraph of $G_{\tau_{2\sigma}}$ induced by E_c^*.

Corollary 2.4.7. *W.h.p.* $|N_c^*(S)| \geq 2|S|$ *for all* $S \subset V(G)$ *with* $|S| \leq \alpha n$.

Proof. We know from Corollary 2.4.4 that w.h.p. at time m every vertex v has $d_c^*(v) \geq 2$. Let $S_2 = S \cap LARGE$, $S_1 = S \setminus S_2$. Then

$$\begin{aligned}|N_c^*(S)| &= |N_c^*(S_1)| + |N_c^*(S_2)| - |N_c^*(S_1) \cap S_2| - |N_c^*(S_2) \cap S_1| - |N_c^*(S_1) \cap N_c^*(S_2)| \\ &\geq |N_c^*(S_1)| + |N_c^*(S_2)| - |S_2| - |N_c^*(S_2) \cap S_1| - |N_c^*(S_1) \cap N_c^*(S_2)|.\end{aligned}$$

Clearly, $|S_2| \leq |S| \leq \alpha n$, and so Lemma 8 gives $|N_c^*(S_2)| \geq 19|S_2|$, w.h.p. Also, recall from Lemma 11 that w.h.p. there are no small structures in G_m and since $SMALL_c \subset SMALL$ w.h.p., this means there aren't any small-c-structures either. In particular,

- No $small_c$ vertices are adjacent and there is no path of length two between $small_c$ vertices which implies that $|N_c^*(S_1)| \geq 2|S_1|$ and $|N_c^*(S_2) \cap S_1| \leq |S_2|$.
- In addition, there is no C_4 containing a $small_c$ vertex, and no path of length 4 between $small_c$ vertices. This means that $|N_c^*(S_1) \cap N_c^*(S_2)| \leq |S_2|$.

We deduce that $|N_c^*(S)| \geq 2|S_1| + 19|S_2| - 3|S_2| \geq 2|S|$. □

Recall Γ_c^* is the subgraph induced by edges of color c that are not in E^+.

Corollary 2.4.8. *W.h.p.* Γ_c^* *is connected for every* $c \in [q]$.

Proof. If $[S, V \setminus S]$ is a cut in Γ_c^* then Corollary 2.4.7 implies that $|S|, |V \setminus S| \geq \alpha n$. Let $F = S \cap Full$. Since $|V \setminus Full| \leq n^{1-\delta} < \frac{\alpha n}{2}$ we see that $|F|, |Full \setminus F| \geq \frac{\alpha n}{2}$. As in **Case 2** of Lemma 8 we show that w.h.p. no such F exists by doing the relevant computation in H_2 with $p_+ \sim \frac{\log n}{4qn}$:

$$\begin{aligned}\mathbb{P}(\exists F) &\leq 2 \sum_{f=\frac{\alpha n}{2}}^{|Full|-\frac{\alpha n}{2}} \binom{|Full|}{f} (1-p_+)^{f(|Full|-f)-t_\epsilon} \\ &\leq n2^n (1-p_+)^{\frac{\alpha n}{2}(n-n^{1-\delta}-\frac{\alpha}{2}n)-o(n^2)} \\ &\leq n2^n e^{-\alpha n \log n/10q} = o(1).\end{aligned}$$

We subtract t_ϵ from $f(|Full| - f)$ because we do not include the first t_ϵ edges in this calculation. This is because $Full$ depends on them. □

2.5 Rotations

We now use E_c^+ to build the Hamiltonian cycles for every color c using Pósa rotations. We let G_c denote the graph induced by the edges of color c. Given a path $P = (x_1, x_2, \ldots, x_k)$ and

an edge $x_ix_k, 2 \le i \le k-2$ we say that the path $P' = (x_1, \ldots, x_i, x_k, \ldots, x_{i+1})$ is obtained from P by a rotation with x_1 as the fixed endpoint.

For a path P in G_c with endpoint a denote by $END(a)$, the set of all endpoints of paths obtainable from P by a sequence of Pósa rotations with a as the fixed endpoint. In this context, Pósa [42] shows that $|N_c(END(a))| < 2|END(a)|$. This is assuming that in the course of executing the rotations, no simple extension of our path is found. It follows from Corollary 2.4.7 that w.h.p. $|END(a)| \ge \alpha n$. For each $b \in END(a)$ there will be a path P_b of the same length as $|P|$ with endpoints a, b. We let $END(b)$ denote the set of all endpoints of paths obtainable from P_b by a sequence of Pósa rotations with b as the fixed endpoint. It also follows from Corollary 2.4.7 that w.h.p. $|END(b)| \ge \alpha n$ for all $b \in END(a)$. Let $END(P) = \{a\} \cup END(a)$.

An edge $u = \{x, y\}$ of color c with $y \in END(x)$ is called a *booster*. Let $P_{x,y}$ be the path of length $|P|$ from x to y implied by $y \in END(x)$. Adding the edge u to $P_{x,y}$ will either create a Hamilton cycle or imply the existence of a path of length $|P|+1$ in G_c, after using Corollary 2.4.8. Indeed, if the cycle C created is not a Hamilton cycle, then the connectivity of Γ_c^* implies that there is an edge $u = xy$ of color c with $x \in V(c)$ and $y \notin V(C)$. Then adding u and removing an edge of C incident to x creates a path of length $|P|+1$.

We start with a longest path in Γ_c^* and let $E_c^+ = \{f_1, f_2, \ldots, f_\ell\}$ where w.h.p. $\ell \ge m_+ = \frac{n \log n}{8q}$, see Claim 2.4.6. A round consists of an attempt to find a longer path than the current one or to close a Hamilton path to a cycle. Suppose we start a round with a path P of length k. We use rotations and construct many paths. If one of these paths has an endpoint with a neighbor outside the path then we add this neighbor to the current path and start a new round with a path of length $k+1$. Here we only use edges not in E_c^+. Failing this we compute $END(P)$ and look for a booster in E_c^+. In the search for boosters we start from f_r assuming that we have already examined $f_1, f_2, \ldots, f_{r-1}$ in previous rounds. Now f_r is chosen uniformly from $(1-o(1))\binom{n}{2}$ pairs and so the probability it is a booster is at least $\beta = (1-o(1))\alpha^2$. It is clear that at most n boosters are needed to create a Hamilton cycle. Adding a booster increases the length of the current path by one, or creates a Hamilton cycle. So the probability we fail to find a Hamilton cycle of color c is at most $\mathbb{P}(Bin(m_+, \beta) \le n) = o(1)$. We can inflate this $o(1)$ by σ to show that w.h.p. we find a Hamilton cycle in each color, completing the proof of Theorem 2.1.1.

2.6 Concluding remarks

In this chapter we studied a very natural variant of the classical problem of the appearance of σ edge disjoint Hamilton cycles in a random graph process. We showed that one can color the edges of the process online so that every color class has a Hamilton cycle exactly at the moment when the underlying graph has σ edge disjoint ones.

The paper [12] shows that at the hitting time $\tau_{2\sigma+1}$ there will w.h.p. be σ edge disjoint Hamilton cycles plus an edge disjoint matching of size $\lfloor n/2 \rfloor$. It is straightforward to extend

this result to the online situation. It should be clear that at time $\tau_{2\sigma+1}$ COL can be used to construct w.h.p. $E_c^*, E_c^+, c = 1, 2, \sigma + 1$ such that $E_c^* \cup E_c^+$ induce Hamiltonian graphs for $1 \leq c \leq \sigma$ and $d_{\sigma+1}^*(v) \geq 1$ for $v \in [n]$. For color $\sigma+1$, we replace the statement of Corollary 2.4.7 by

$$\text{W.h.p. } |N_{\sigma+1}^*(S)| \geq |S| \text{ for all } S \subset V(G) \text{ with } |S| \leq \alpha n. \tag{2.15}$$

We then replace rotations by alternating paths, using $E_{\sigma+1}^+$ as boosters. The details are as described in Chapter 6 of [26]. In outline, let $G = (V, E)$ be a graph without a matching of size $\lfloor |V(G)|/2 \rfloor$. For $v \in V$ such that v is isolated by some maximum matching, let

$$A(v) = \{w \in V : w \neq v \text{ and } \exists \text{ a maximum matching of } G \text{ that isolates } v \text{ and } w\}.$$

We use the following lemma

Lemma 9. *Let G be a graph without a matching of size $\lfloor |V(G)|/2 \rfloor$. Let M be a maximum matching of G. If $v \in V$ and $A(v) \neq \emptyset$ then $|N_G(A(v))| < |A(v)|$.*

We start with a maximum matching M of $\Gamma_{\sigma+1}^*$. Suppose that v is not covered by M. Using (2.15), we see that w.h.p. $|A(v)| \geq \alpha n$. Further, if $u \in A(v)$ and $uv \in E_{\sigma+1}^+$ then adding this edge gives a larger matching. Also, because u is isolated by a maximum matching, there is a corresponding set A_u of size at least αn such if $w \in A_u$ and $uw \in E_{\sigma+1}^+$ then we can find a larger matching. Therefore we have $\Omega(n^2)$ boosters and the proof is similar to that for Hamilton cycles.

There are several related problems which can likely be treated using our approach. One potential application for our technique is to show that for any fixed positive integer k and any decomposition $k = k_1 + ... + k_s$ into the sum of s positive integers, there is an online algorithm, coloring the edges of a random graph process in s colors so that exactly at the hitting time τ_k the i-th color forms a k_i-connected spanning graph for $i = 1, \ldots, s$. In general, one can generate many more interesting problems by considering the online Ramsey version of other results in the theory of random graphs.

Chapter 3

Directed Hamilton Cycles

3.1 Introduction

Let $\vec{K}_n$ be the complete directed graph on n vertices. We let $(e_1, e_2, ..., e_{n(n-1)})$ be a uniformly random permutation of the edges of $\vec{K}_n$ and consider the random process of digraphs $D_1, D_2, ..., D_{n(n-1)}$ defined by $D_m = (V_n, E_m)$ with $E_m = (e_1, ..., e_m)$ for $m \in [n(n-1)]$. This is a directed analogue of the celebrated Erdős-Rényi random graph process [19], in which the edges of the *undirected* complete graph K_n are ordered uniformly at random, similarly yielding a random process of graphs $G_1, G_2, \ldots, G_{n(n-1)/2} = K_n$. Graph-theoretic properties of D_m and G_m are said to hold "with high probability" (w.h.p.) if they occur with probability $1 - o(1)$ as $n \to \infty$, where m is allowed to be a random variable depending on n.

A *Hamilton cycle* is a (directed) cycle passing through all n vertices exactly once. When a graph or digraph contains such a cycle, we say it is *Hamiltonian*. The study of Hamilton cycles is fundamental to graph theory, including in the random setting. For a digraph to contain a Hamilton cycle it certainly requires each vertex to have at least 1 in-edge and 1 out-edge, but quite remarkably, this is almost always sufficient for the random graphs D_m. Specifically, for a fixed q, let D_{τ_q} denote the first digraph in this random process with both minimum in-degree and out-degree $\geq q$. In [27], Frieze showed that w.h.p. D_{τ_1} is Hamiltonian yielding a hitting-time strengthening of McDiarmid [41] and a directed version of the classical result due to Bollobás [9] and Ajtai, Komlós and Szemerédi [3]. The latter two papers independently proved that w.h.p. the first G_m in the undirected random graph process with minimum degree $\delta(G_m) \geq 2$ is Hamiltonian, thus bringing to fruition the work built up by Komlós and Szemerédi [34], Korshunov [35] and Pósa [42] previously.

The undirected version was strengthened [12] by Frieze and Bollobás to additional Hamilton cycles thus: let $q = O(1)$ be fixed. Let $G_{\tau'_{2q}}$ be the first random graph in the undirected process with $\delta(G_{\tau'_{2q}}) = 2q$ (here τ_q and τ'_q distinguish the directed and undirected hitting times respectively). Then w.h.p. $G_{\tau'_{2q}}$ has a q-edge-coloring with a Hamilton cycle in every color. In fact, results for $q \to \infty$ with $n \to \infty$ have been established in all cases thanks to

extensive work completed by Knox, Kühn and Osthus [33] and Krivelevich and Samotij [38].

In these papers, it appeared that the minimum degree conditions were still the most binding aspects of the proofs, suggesting stronger results could be obtained if corresponding minimum degree conditions are met. Indeed, Krivelevich, Lubetzky and Sudakov [37] took advantage of the Achlioptas process with parameter $K = o(\log n)$ to build a Hamilton cycle using w.h.p. only $(1 + o(1))\frac{\tau_2'}{K}$ edges. In this process, at each time step, K random new edges are presented, out of which one is added to the current graph, thereby allowing a bias towards low-degree vertices when necessary. For Theorem 2.1.1 of Chapter 2, we presented an algorithm coloring the edges $(e_1, e_2, ..., e_{n(n-1)/2})$ as they appeared, with $q = O(1)$ colors, such that w.h.p. $G_{\tau_{2q}'}$ contains a monochromatic Hamilton cycle of every color. The on-line nature of this strengthening of the classical result is of importance, because the color of each new random edge e_m *cannot depend on the location of the edges appearing thereafter.*

We now consider the analogous scenario in the directed random graph process. Here, the edges of the random permutation $(e_1, e_2, ..., e_{n(n-1)})$ of $\vec{K}_n$ are revealed one by one. As soon as an edge is revealed it has to be colored irrevocably with one of $q = O(1)$ colors. We prove the following:

Theorem 1. *There exists an on-line $[q]$-edge-coloring algorithm for $D_1, \ldots, D_{n(n-1)}$ such that w.h.p. D_{τ_q} has q monochromatic Hamilton cycles, one in every color in $[q]$.*

In order to prove Theorem 1 we present a coloring algorithm which we name COL. Thereafter we split the proof into two parts. In the first part we prove that each color class c of D_{τ_q} given by $DCOL$ satisfies the minimum degree condition necessary for Hamiltonicity. In the second part (drawing our proof strategy from [27]) we fix $c \in [q]$ and show w.h.p. D_{τ_q} has a monochromatic Hamilton cycle in color c. To do so we end up giving a reduction to the following more general Lemma.

Lemma 10. *Let $F, H, D_{n,p}$ be digraphs on the same vertex set of size n such that:*

i) F is a 1-factor consisting of $O(\log n)$ directed cycles,

ii) H has maximum in/out-degree $O(\log n)$,

iii) $D_{n,p}$ is the random digraph where every edge appears independently with probability $p = \Omega(\frac{\log n}{n})$.

Then w.h.p. there is a Hamilton cycle spanned by $E(F) \cup \big(E(D_{n,p}) \setminus E(H)\big)$.

Throughout this chapter we use the well-known result (see for example [26]) that w.h.p.

$$n \log n + n(q-1) \log \log n - \omega \leq \tau_q, \tau_q' \leq n \log n + n(q-1) \log \log n + \omega$$

for any $\omega = \omega(n)$ which tends to infinity as n tends to infinity.

3.2 The Colouring Algorithm $DCOL$

The coloring algorithm $DCOL$, given shortly, will color greedily arcs that are incident to vertices that "do not see all the colors yet". These vertices are the most dangerous, as indeed some will only have q out-arcs in D_{τ_q}, accordingly needing *exactly* 1 of each color. We formalize these most needy of vertices by means of the notation in the following subsection, to guide our description of the algorithm $DCOL$. Note that the notation given below will be used repeatedly throughout this chapter.

3.2.1 Some notation

Notation 3.2.1. "By/at time t" is taken to mean "after t edges have been revealed", that is, with respect to D_t. We also write τ for τ_q.

Definition 3.2.2. For $v \in V_n$, $c \in [q]$ and $t \in \{0, 1, ..., \tau\}$, we set $d_t^+(v, c)$ (and $d_t^-(v, c)$ resp.) to equal the numbers of arcs with out-(in- resp.) vertex v, that have been revealed by time t and have been assigned color c by the algorithm DCOL. Also write $d_t^+(v)$ ($d_t^-(v)$ resp) for the total number of out- (in-) arcs from v by time t. Hence $d_t^+(v) = \sum_{c\in[q]} d_t^+(v, c)$ at any time t. For the final in/out-degrees we write $d^-(v) := d_\tau^-(v)$ and $d^+(v) := d_\tau^+(v)$.

Definition 3.2.3. For $v \in V_n$ and $t \in \{0, 1, ..., \tau\}$ we set $C_v^+(t)$:=$\{c \in [q] : d_t^+(v, c) = 0\}$ (i.e the colors that at time t are missing from the out-arcs of v). Similarly set $C_u^-(t)$:=$\{c \in [q] : d_t^-(v, c) = 0\}$.

Definition 3.2.4. For $t \in \{0, 1, ..., \tau\}$ we set $FULL_t^+$:=$\{v \in V_n : C_v^+(t) = \emptyset\}$ (i.e. the set of vertices that at time t have out degree in each color at least one). Similarly define $FULL_t^-$. We certainly want both $FULL_\tau^+, FULL_\tau^-$ to contain all of V_n in the end.

3.2.2 Algorithm $DCOL$

Algorithm ColorGreedy(u, v, t) will be called in multiple places during the algorithm $DCOL$, hence is given beforehand.

Algorithm 1 ColorGreedy(u, v, t)

if $u \notin FULL_{t-1}^+$ *or* $v \notin FULL_{t-1}^-$ **then**
 | color arc uv by a color chosen uniformly at random from $C_u^+(t-1) \cup C_v^-(t-1)$.
else
 | color arc uv by a color chosen uniformly at random from $[q]$.
end

For $i \in \{0,1,2,3\}$ we also set $m_i = i \cdot e^{-q \cdot 10^4} n \log n$, marking out 3 small but positive fractions of the (expected) number of edges τ, and $p_i = \frac{m_i}{n(n-1)}$.

Algorithm 2 DCOL

for $t = 1, ..., m_1$ **do**
  let $e_t = uv$
  Execute ColorGreedy(u, v, t).
end
For $v \in V_n$ set $c^+(v) = 1$, $c^-(v) = 1$.
for $t = m_1 + 1, m_1 + 2, ..., m_2$ **do**
  let $e_t = uv$
  if $u \notin FULL^+_{t-1}$ *or* $v \notin FULL^-_{t-1}$ **then**
    Execute ColorGreedy(u, v, t).
  else
    Color the arc uv by the color c satisfying $c \equiv c^+(u) \mod q$,
    $c^+(u) \leftarrow c^+(u) + 1$.
  end
end
for $t = m_2 + 1, m_2 + 2, ..., m_3$ **do**
  let $e_t = uv$
  if $u \notin FULL^+_{t-1}$ *or* $v \notin FULL^-_{t-1}$ **then**
    Execute ColorGreedy(u, v, t).
  else
    Color the arc uv by the color c satisfying $c \equiv c^-(v) \mod q$,
    $c^-(v) \leftarrow c^-(v) + 1$.
  end
end
For $i \in \{1,2,3\}, * \in \{+,-\}$ set $B^*_i := \{v \in V_n : d^*_{m_i}(v) - d^*_{m_{i-1}} \leq \epsilon \log n\}$, where $\epsilon = e^{-q \cdot 10^6}$.
Furthermore set $BAD := B^+_1 \cup B^-_1 \cup B^+_2 \cup B^-_3$ and $E' := \emptyset$.
for $t = m_3 + 1, ..., \tau$ **do**
  let $e_t = uv$
  if $u \notin FULL^+_{t-1}$ *or* $v \notin FULL^-_{t-1}$ **then**
    Execute ColorGreedy(u, v, t)
  else if $u \in BAD$ *or* $v \in BAD$ **then**
    Color the arc uv by a color c that minimizes
    $d^+_t(u, c)\mathbb{I}(u \in BAD) + d^-_t(v, c)\mathbb{I}(v \in BAD)$. If there is more than one such color then choose one from them uniformly at random.
  else
    Execute ColorGreedy(u, v, t).
    Add the arc uv to E'.
  end
end

Remark 3.2.5. Suppose at some time t that $e_t = uv$ and $C_u^+(t-1) \cup C_v^-(t-1) \neq \emptyset$, i.e. $u \notin FULL_{t-1}^+$ is still missing an out-edge color or $v \notin FULL_{t-1}^-$ is still missing an in-edge color. Then any color from $C_u^+(t-1) \cup C_v^-(t-1)$ has probability at least $\frac{1}{q}$ to be chosen to color uv.

Remark 3.2.6. The second priority (after the vertices needing to be greedy) is to build the 1-factor F in each color needed to power Lemma 10, for which we aim to have as many vertices with at least a prescribed out-degree as possible (in fact, 6 will do). The cycling with c^+ and c^- between edge colors during times $(m_1, m_2]$ and $(m_2, m_3]$ will ensure as many of the $FULL$ vertices as possible receive an ample balance of edges in each color. The few exceptions are confined to BAD and forced to balance their colors for the remainder of the process. Meanwhile, the arcs in E' enjoy full randomness, and can be used to build the desired Hamilton cycles using classical techniques.

3.3 Structural results

Recall the following relations between $D_{n,m}$ and $D_{n,p}$ (see [26]). Let Q be any property of $D_{n,m}$ for some m, $0 \leq m \leq n(n-1)$ and let $p = \frac{m}{n(n-1)}$ then,

$$\mathbb{P}(D_{n,m} \text{ has } Q) \leq 10\sqrt{m}\mathbb{P}(D_{n,p} \text{ has } Q).$$

Moreover if Q is a monotone increasing property i.e. it is preserved under edge addition or monotone decreasing property i.e. it is preserved under edge deletion, then we have

$$\mathbb{P}(D_{n,m} \text{ has } Q) \leq 3\mathbb{P}(D_{n,p} \text{ has } Q).$$

For $p \in [0,1]$ we denote by $Bin(k,p)$ the random variable following the Binomial distribution with k objects each appearing with probability p. Also, we will make use of the Chernoff bounds (see [31]): namely, if X is a $Bin(k,p)$ random variable with mean $\mu = np$ then for any $\epsilon > 0$ we have

$$Pr[X \leq (1-\epsilon)\mu] \leq e^{-\frac{\epsilon^2\mu}{2}},$$
$$Pr[X \geq (1+\epsilon)\mu] \leq e^{-\frac{\epsilon^2\mu}{2+\epsilon}}.$$

Finally for the rest of the chapter we let

$$p_\ell = \frac{\log n + (q-1)\log\log n - \omega(n)}{n}, \qquad m_\ell = n(n-1)p_\ell,$$

and

$$p_u = \frac{\log n + (q-1)\log\log n + \omega(n)}{n}, \qquad m_u = n(n-1)p_u,$$

where $\omega(n) = \frac{1}{2}\log\log\log n$. Recall that w.h.p. D_{n,m_ℓ} has zero vertices of in- or out- degree less than $q-1$. In addition w.h.p. $m_\ell \leq \tau \leq m_u$.

Lemma 11. *W.h.p. for $k \in \big[q-1, \frac{3\log n}{\log\log n}\big]$, D_{n,m_ℓ} has at most $v_k := \frac{e^{2\omega(n)}(\log n)^{k-q+1}}{(k-1)!}$ vertices of in-degree at most k. Hence, the same is true for vertices of in-degree exactly k, and similarly for out-degree k.*

Proof. By taking a union bound and using (2) for the first inequality, we get

$$\begin{aligned}
&\mathbb{P}(D_{n,m_\ell} \text{ has more than } v_k \text{ vertices of in-degree at most } k)\\
&\le \binom{n}{v_k}\left[3\sum_{j=0}^{j=k}\binom{n-1}{j}(1-p_\ell)^{n-j-1}p_\ell^j\right]^{v_k} \le \left(\frac{en}{v_k}\right)^{v_k}\left[3(k+1)\binom{n-1}{k}(1-p_\ell)^{n-k-1}p_\ell^k\right]^{v_k}\\
&\le \left[\frac{en}{v_k}\frac{3(k+1)n^k}{k!}e^{-\log n-(q-1)\log\log n+\omega(n)+o(1)}\left(\frac{\log n+(q-1)\log n\log n-\omega(n)}{n}\right)^k\right]^{v_k}\\
&\le \left[e^{-\omega(n)+O(1)}\left(1+\frac{q\log\log n}{\log n}\right)^k\right]^{v_k} \le \left[e^{-\omega(n)+O(1)+\frac{q\log\log n}{\log n}k}\right]^{v_k} \le e^{-\frac{\omega(n)v_k}{2}}.
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{P}\Bigg(\text{for some } k \in \left[q-1, \frac{3\log n}{\log\log n}\right] \text{ there are } > v_k \text{ vertices of out-degree k in } D_{n,m_\ell}\Bigg)\\
&\qquad \le \sum_{k=q-1}^{\frac{3\log n}{\log\log n}} e^{-\frac{\omega(n)v_k}{2}} = \sum_{k=q-1}^{\frac{3\log n}{\log\log n}} \left(e^{-\frac{1}{4}\log\log\log n}\right)^{v_k} = o(1).
\end{aligned}$$

□

Definition 3.3.1. For $u, v \in V_n$ let the undirected distance from u to v at time t, denoted by $d'_t(u,v)$, be the distance from u to v in the graph that is obtained from D_t when we ignore the orientations of the edges.

Definition 3.3.2. Let $SMALL := \{v \in V : d^+_\tau(v) \le \frac{\log n}{100} \text{ or } d^-_\tau(v) \le \frac{\log n}{100}\}$. Since we expect $\tau \ge n\log n$, $SMALL$ consists of vertices with significantly smaller degree than their expected value.

Lemma 12. *W.h.p. for every $v, w \in SMALL, d'_\tau(v,w) \ge 2$.*

Proof. We weaker the definition of $SMALL$ so that it suffices to do the computation in D_{m_u}. Specifically, set $SMALL'$:=$\{v \in V : d^+_{m_u}(v) \text{ or } d^-_{m_u}(v) \le \frac{1}{100}\log n + 2\omega(n)\}$. (1) gives us

$$\begin{aligned}
&\mathbb{P}\big(v, w \in SMALL' \text{ and } d'_{m_u}(v,w) \le 2\big)\\
&\qquad \le 10\sqrt{m_u}\sum_{k=1,2}\binom{n-2}{k-1}(2p_u-p_u^2)^k\left[2\mathbb{P}\left(Bin(n-1-k,p_u) \le \frac{\log n}{100}+2\omega(n)-1\right)\right]^2\\
&\qquad \le \frac{200\sqrt{n}\log^{2.5}n}{n}\left[\exp\left(-(1-o(1))\log n\left(\frac{1}{100}\log\frac{1}{100}+\frac{99}{100}\right)\right)\right]^2 = o(n^{-2.3}).
\end{aligned}$$

At the second inequality we used that $\mathbb{P}(Bin(\lambda/p,p) \le \lambda - t) \le \exp\{-\lambda[(1+x)\log(1+x)-x]\}$ (see [31]), with $x = -\frac{t}{\lambda} \sim -\frac{99}{100}$ for $\lambda = (n-1-k)p_u \sim \log n, t = \lambda - \frac{\log n}{100} - 2\omega(n)$ here. In the

event $m_\ell \le \tau \le m_u$, as D_τ precedes D_{m_u}, we have that $E_\tau \subseteq E_{m_u}$ and $|E_{m_u} \backslash E_\tau| \le 2\omega(n)$. Furthermore if $d'_\tau(v,w) \le 2$ then $d'_{m_u}(v,w) \le 2$. Therefore $m_\ell \le \tau \le m_u$ implies that $SMALL \subseteq SMALL'$. Hence,

$$\mathbb{P}\Big(\exists v,w \in SMALL \text{ such that } d'_\tau(v,w) \le 2\Big) \le \binom{n}{2} o(n^{-2.3}) + \mathbb{P}\Big(\tau \notin [m_\ell, m_u]\Big) = o(1). \quad \square$$

Notation 3.3.3. For a digraph D denote by $\Delta^+(D)$ and $\Delta^-(D)$ its maximum out- and in-degree respectively. Futhermore set $\Delta(D) = \max\{\Delta^+(D), \Delta^-(D)\}$.

Lemma 13. *W.h.p.* $\Delta(D_\tau) \le 12\log n$.

Proof. We implicitly condition on the event $\{\tau \le m_u\}$. Using (2),

$$\begin{aligned}\mathbb{P}\Big(\Delta^+(D_\tau) \text{ or } \Delta^-(D_\tau) \ge 12\log n\Big) &\le 3\cdot 2n\binom{n-1}{12\log n} p_u^{12\log n} \\ &\le 6n\left(\frac{en}{12\log n}\right)^{12\log n} p_u^{12\log n} \le 6n\left(\frac{en}{12\log n}\cdot\frac{2\log n}{n}\right)^{12\log n} = o(1). \quad \square\end{aligned}$$

Lemma 14. *W.h.p.* $\Delta(D_{m_1}) \le \frac{\log n}{10^3 q}$.

Proof. Recall $p_1 = \frac{m_1}{n(n-1)} = \frac{e^{-q\cdot 10^4}\log n}{n-1}$. Then (2) gives us that

$$\begin{aligned}\mathbb{P}\left(\Delta^+(D_{m_1}) \text{ or } \Delta^-(D_{m_1}) \ge \frac{\log n}{10^3 q}\right) &\le 3\cdot 2n\binom{n-1}{\frac{\log n}{10^3 q}} p_1^{\frac{\log n}{10^3 q}} \le 6n\left(\frac{10^3 qe(n-1)}{\log n}\right)^{\frac{\log n}{10^3 q}} p_1^{\frac{\log n}{10^3 q}} \\ &\le 6n\left(10^3 q e^{-q\cdot 10^4+1}\right)^{\frac{\log n}{10^3 q}} = o(1). \quad \square\end{aligned}$$

3.4 Minimum degree 1 in color c

Theorem 2. *W.h.p. DCOL succeeds in assigning colors to the arcs so that* $\forall c \in [q]$ *and* $\forall v \in V_n$ *we have* $d^+_\tau(v,c)$, $d^-_\tau(v,c) \ge 1$.

We will approach this theorem by conditioning on the final digraph D_τ (in particular, on Lemmas 11 and 12) and analysing the randomness of the edges' order and color. By symmetry, it suffices to prove the out-degree part. The proof will follow from Lemmas 16, 17 given below.

For most of this section, at least until Lemma 17, $v \in V_n$ will be arbitrary but (crucially) fixed. Denote by $N^+(v)$ the out-neighbours of v in D_τ and set $N^+_L(v) := N^+(v)\backslash SMALL_\tau$- we aim for these larger neighbours to provide v with the colors it needs, and thankfully,

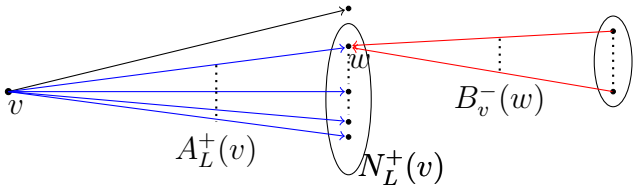


Figure 3.1: Arcs in $A_L^+(v)$ and in $B_u^-(w)$ are in blue and red respectively.

Lemma 12 ensures ≤ 1 neighbour was in $SMALL$. Furthermore let $A_L^+(v)$ be the set of arcs arising from $N_L^+(v)$ (i.e. $A_L^+(v) := \{vw \in E_\tau : w \in N_L^+(v)\}$). For $w \in N_L^+(v)$ we fix a set $B_v^-(w)$ of $\frac{\log n}{100} - 1$ arcs in $\big(V_n \backslash \{v, w\}\big) \times \{w\}$. Finally we let $A_v^-(w) := B_v^-(w) \cup \{vw\}$.
We will only need to analyse the algorithm's effect on $\bigcup_w A_v^-(w)$ to show v is unlikely to obtain all the colors it needs. For this analysis, we couple the algorithm as follows. Let $D_\tau^{(1)}$ and $D_\tau^{(2)}$ be two copies of D_τ colored in parallel according to algorithm $DCOL1(v)$ given below. $D_\tau^{(2)}$ will mimic $DCOL$. Meanwhile, $D_\tau^{(1)}$ will be strictly worse (for v's satisfaction), but will color $\bigcup_w B_v^-(w)$ fully randomly, and thus will be easier to analyse.

Notation 3.4.1. For $i \in [2]$ we extend the notation $C_v^+(t)$, $C_v^-(t)$, $FULL_t^+$, $FULL_t^-$, BAD to $C_{i,v}^+(t)$, $C_{i,v}^-(t)$, $FULL_{i,t}^+$, $FULL_{i,t}^-$ and BAD_i for the corresponding sets in $D_\tau^{(i)}$.

Algorithm 3 DCOL1(v)

for $t = 1, ..., \tau$ **do**
  let $e_t = xy$
  if $e_t \in \bigcup_{w \in N_L^+(v)} B_v^-(w)$ **then**
    choose a color c from $[q]$ uniformly at random
    if $c \in C_{2,x}^+(t-1) \cup C_{2,y}^-(t-1)$ **then**
      color e_t in both $D_\tau^{(1)}$, $D_\tau^{(2)}$ with color c.
    else
      color e_t in $D_\tau^{(1)}$ with color c,
      to color e_t in $D_\tau^{(2)}$ execute step t of $DCOL$, [1]
    end
  else
    to color e_t in $D_\tau^{(2)}$ execute step t of $DCOL$. [1]
    color e_t in $D_\tau^{(1)}$ by the same color as in $D_\tau^{(2)}$.
  end
end

[1] Here we suppose that we run $DCOL$. Our current arcs $e_1, ..., e_{(t-1)}$ have the colors that have been assigned by $DCOL1(v)$ to the corresponding arcs in $D_\tau^{(2)}$. We use $FULL_{2,t}^+$, $FULL_{2,t}^-$ and BAD_2 in place of $FULL_t^+$, $FULL_t^-$ and BAD respectively.

Remark 3.4.2. The colorings of $D_\tau^{(2)}$ and D_τ have the same distribution.

Remark 3.4.3. For every $t \in [\tau]$ and $w \in N_L^+(v)$ since the algorithm may color an arc $e_t = xw$ in $D_\tau^{(1)}$ and in $D_\tau^{(2)}$ with distinct colors c and c' respectively only in the case where $c \notin C_{2,x}^+(t-1) \cup C_{2,w}^-(t-1)$ (i.e $c \notin C_{2,w}^-(t-1)$) we have $C_{2,w}^-(t) \subseteq C_{1,w}^-(t)$.

Definition 3.4.4. Fot $t \in [\tau]$ we say that $e_t \in A^+(v)$ contributes to the coloring of v (or just contributes to v) in $D_\tau^{(1)}$ if *either* $C_{1,v}^+(t-1) = \emptyset$ *or* e_t gets a color in $C_{1,v}^+(t-1)$.

Lemma 15. *Once q arcs have contributed to the coloring of v in $D_\tau^{(1)}$ we have that in $D_\tau^{(2)}$, v has out-degree at least one in each color.*

Proof. Follows directly from Definition 3.4.4 and Remark 3.4.3. □

The strength of Lemma 15 is that it allows us to do the desired computations in $D_\tau^{(1)}$, for Lemmas 16 and 17.

Lemma 16. *For any $v \in V_n$, if we run the corresponding coloring algorithm $D_\tau^{(1)}$:*

$$\mathbb{P}\big(\textit{less than } q \textit{ arcs contribute to the coloring of } v \textit{ in } D_{m_\ell}^{(1)}\big) \leq \binom{d^+(v)-1}{q-1}\left(\frac{100q^{q+1}}{\log n}\right)^{d^+(v)-q}.$$

Before proceeding to the proof of Lemma 16 we introduce the following two functions.

Definition 3.4.5. For $e \in E_\tau$ define the bijection $h : E_\tau \to [\tau]$ where $h(e) = k$ means $e = e_k$, i.e e was the kth arc to be revealed. Thus, for example, $FULL_{1,h(vw)}^- = FULL_{1,t'}^-$ where $e_{t'} = vw$.

Definition 3.4.6. For $w \in N_L^+(v)$ define the bijection $g_{v,w} : A_v^-(w) \to \left[\frac{\log n}{100}\right]$ where $g_{v,w}(xw) = k$ means xw is the kth arc that was revealed out of all the arcs in $A_v^-(w)$.

Also we define the following events.

Definition 3.4.7. For $w \in N_L^+(v)$ set $F(w)$ to be the event that in $D_\tau^{(1)}$ $\nexists \ell \in \mathbb{Z}_{\geq 0}$ s.t. $\ell q + q < g_{v,w}(vw)$ and $g_{v,w}^{-1}(\ell q + 1), ..., g_{v,w}^{-1}(\ell q + q)$ are colored by q distinct colors.

Remark 3.4.8. For every $w \in N_L^+(v)$, the event $\{w \notin FULL_{1,h(vw)}^-\} \subseteq F(w)$.

Indeed, for any $\ell \in \mathbb{Z}_{\geq 0}$ such that $\ell q + q < g_{v,w}(vw)$ the arcs $g_{v,w}^{-1}(\ell q + 1), ..., g_{v,w}^{-1}(\ell q + q)$ precede vw. So if they were colored differently, we would have $w \in FULL_{1,h(vw)}^-$, which is the contrapositive.

Remark 3.4.9. The events $\{F(w) : w \in N_L^+(v)\}$ are independent.

Indeed, for $w \in N_L^+(v), \mathbb{P}(F(w))$ depends only on the relative time $g_{v,w}(vw)$ of vw among in-edges of w. That is because the colors that *DCOL1(v)* assigns to the edges, $g_{v,w}^{-1}(1), g_{v,w}^{-1}(2), ...,$ $g_{v,w}^{-1}\big(g_{v,w}(vw)-1\big)$, preceding vw are chosen independently and uniformly at random from $[q]$. Thus in showing the independence of $\{F(w)\}$ it suffices to note that the values $\{g_{v,w}(vw) : w \in N_L^+(v)\}$ are independent, and this follows from the sets $A_v^-(w)$ being disjoint.

Proof of Lemma 16: For $w \in N_L^+(v)$,

$$\begin{aligned}\mathbb{P}\big(F(w)\big) &= \sum_{k=1}^{\frac{\log n}{100}} \mathbb{P}\big(\{g_{v,w}(vw)=k\} \wedge F(w)\big) = \sum_{k=1}^{\frac{\log n}{100}} \mathbb{P}\big(g_{v,w}(vw)=k\big)\mathbb{P}\big(F(w)|g_{v,w}(vw)=k\big) \\ &\leq \sum_{k=1}^{\frac{\log n}{100}} \frac{100}{\log n} \prod_{l=1}^{\lfloor (k/q)-1 \rfloor} \left(1-\frac{1}{q^q}\right) \leq \frac{100}{\log n} \sum_{j \in \mathbb{Z}_{\geq 0}} q\left(1-\frac{1}{q^q}\right)^j \leq \frac{100}{\log n} q^{q+1}.\end{aligned} \tag{3.1}$$

Hence,

$$\begin{aligned}&\mathbb{P}\big(\text{less than } q \text{ arcs contribute to the coloring of } v \text{ in } D_\tau^{(1)}\ \big) \\ &\leq \mathbb{P}\left(\left|\left\{w \in N_L^+(v) : w \notin FULL^-_{1,g_{v,w}^{-1}(vw)}\right\}\right| \geq d^+(v)-q\right) \\ &\leq \mathbb{P}\left(\left|\left\{w \in N_L^+(v) : \text{event } F(w) \text{ occurs }\right\}\right| \geq d^+(v)-q\right) \\ &\leq \mathbb{P}\left(Bin\left(d^+(v)-1, \frac{100q^{q+1}}{\log n}\right) \geq d^+(v)-q\right) \leq \binom{d^+(v)-1}{q-1}\left(\frac{100q^{q+1}}{\log n}\right)^{d^+(v)-q}. \quad \square\end{aligned}$$

The second inequality follows from Remark 3.4.8. The last inequality follows from the independence of the events $\{F(w)\}$, the fact that $\left|N_L^+(v)\right| \geq d^+(v)-1$ (see Lemma 12) and (3.1).

Remark 3.4.10. The two basic ingredients that are used in the proof of Lemma 16 as well as in Lemma 17 are the following: First, for $w \in N_L^+(v)$ the sets $B_v^-(w)$ are disjoint and of size $\Omega(\log n)$. Second, in $D_\tau^{(1)}$ for every $w \in N_L^+(v)$ the arcs in $B_v^-(w)$ are colored independently and uniformly at random. The disjointness of the sets $B_v^-(w)$ implied the independence of the events $F(w)$ while the fact their size is $\Omega(\log n)$ leads to the desired probability being sufficiently small.

The following remark will be used later in the proof of Lemma 22:

Remark 3.4.11. We could reproduce the above lemma with different parameters and similar definitions. That is we could use m_1 in place of τ, $N_{m_1}^+(v)$ to be the neighbours of v in D_{m_1} and for $w \in N_{m_1}^+(v)$ $B_{m_1,v}^-(w)$ to be a set of arcs in E_{m_1} from $V_n \backslash \{v,w\}$ to w of size $\gamma \log n$ where γ is some positive constant. In this case for every $v \in V_c$ such that the condition $|\{w \in V_n : w \in N^+(v), h(vw) < m_1 \text{ and } d_{m_1}^+(w) \leq \gamma \log n\}| \leq k$ (in place of Lemma 12) holds, using the same methodology, we could prove that

$$\mathbb{P}\big(\text{less than } q \text{ arcs contribute to } v \text{ in } D_{m_1}^{(1)}(v)\big) \leq \binom{d_{m_1}^+(v)-k}{q-1}\left(\frac{q^{q+1}}{\gamma \log n}\right)^{\left(d_{m_1}^+(v)-k\right)-(q-1)}.$$

Hence, setting $d = \min\{d^+_{m_1}(v), d^-_{m_1}(v)\}$, we have

$$\mathbb{P}\big(v \notin FULL^+_{m_1} \cap FULL^-_{m_1}\big) \leq 2\binom{d-k}{q-1}\left(\frac{q^{q+1}}{\gamma \log n}\right)^{(d-k)-(q-1)}.$$

The bound provided by Lemma 16 is not strong enough for vertices of small out-degree. However, it can be improved by considering some extra information, provided by Lemma 17. Suppose $e_{\tau_q} = (v^*, w^*)$. Since e_{τ_q} is the last arc of our process we have that either $d^+(v^*) = q$ or $d^-(w^*) = q$. In the case that $d^+(v^*) = q$ we handle v^* separately. Otherwise $d^-(w^*) = q$ and Lemma 12 implies that $d^+(v^*) > \frac{\log n}{100}$. We may assume that $d^+(v^*) = q$ and we deal with v^* separately later.

Lemma 17. *Let $v \in V_n \setminus v^*$ satisfy $q \leq d^+(v) \leq \log\log n$. Then the probability that fewer than q arcs contribute to the coloring of v in $D^{(1)}_\tau$ is bounded above by*

$$\frac{101(\log\log n)^5}{\log n}\binom{d^+(v)-1}{q-2}\left(\frac{101q^{q+1}}{\log n}\right)^{d^+(v)-q+1}.$$

In addition to the $\{g_{v,w}\}$ keeping track of the (random) relative timings of edges within each $A^-_v(w)$, we also care about the relative timings of edges within our entire subgraph $\bigcup_w A^-_v(w)$ and also within our most crucial edges $A^+_L(v)$ that we hope will contribute to v. We define the following two functions accordingly:

Definition 3.4.12. For each $v \in V_n$, let $g_v : A^+_L(v) \to \big[|A^+_L(v)|\big]$ map $vw \mapsto k$ whenever vw is the kth arc revealed among $A^+_L(v)$. Similarly define $h_v : \bigcup_{w \in N^+_L(v)} A^-_v(w) \to \big[\frac{\log n}{100} \cdot |A^+_L(v)|\big]$.

Observe that the maps $h_v(\cdot), g_v(\cdot)$ are also bijections.

Proof of Lemma 16: Our strategy is as thus. Most of the time, we expect that none of the crucial edges in $A^+_L(v)$ appear before some time $r \ll \frac{\log n}{100}$, by which point we also expect that all $w \in N^+_L(v)$ have received a reasonable collection $1 \ll r_\ell \ll r$ of their own edges from *other* vertices. It is unlikely that either of these heuristics fail (see bounds on $\mathbb{P}(A)$ and $\mathbb{P}(B)$ in Cases 1 and 2 below), and when they are correct (Case 3), all the w's become measurably more likely to have become $FULL$ by the time edge vw appears. Specifically, with $r_\ell = q^q \log\log n$ and $r = (\log\log n)^5$ we define the events A and B:

- Let A be the event $\{h_v(g_v^{-1}(1)) \leq r\}$; i.e. the first arc of $A^+_L(v)$ precedes the $(r+1)$st of $\bigcup_{w \in N^+_L(v)} A^-_v(w)$.
- Let B be the event $\{\exists w \in N^+_L(v) : h_v(g^{-1}_{v,w}(r_\ell)) > r+1\}$; i.e. for some $w \in N^+_L(v)$, less than r_ℓ arcs in $A^-_v(w)$ are revealed before the $(r+1)$st arc of $\bigcup_{w \in N^+_L(v)} A^-_v(w)$.

We condition on whether A, $A^c \cap B$, or $A^c \cap B^c$ occurs. In each case we use the same methodology as in Lemma 16 to bound the desired probability. Observe that Lemma 12 implies, as $d^+_\tau(v) \leq \log\log n$, that v has no out-neighbour in $SMALL_\tau$, hence $N^+(v) = N^+_L(v)$. Furthermore note that in any of the events A, $A^c \cap B$ and $A^c \cap B^c$ the first arc that appears with out-vertex v contributes to the colouring of v. Since $N^+(v) = N^+_L(v)$ that arc belongs to $A^+_L(v)$.

• Case 1: A occurs. We describe the possible offending sequences leading up to the early first edge of $A^+_L(v)$ as follows.

Set $\mathcal{E}_1 = \left\{(f_1, ...f_s) \in \left(\bigcup_{w\in A^+_L(v)} B^-_v(w)\right)^{s-1} \times A^+_L(v) : s \leq r \text{ and } f_1, ..., f_s \text{ are distinct}\right\}$.
For $E = (f_1, ...f_s) \in \mathcal{E}_1$ we set $f_E := f_s$ and define A_E to be the event where both:

- f_s is the first arc to be revealed from $A_L(v)$, and
- $f_1, ..., f_{s-1}$ are the only arcs in $\bigcup_{w\in A^+_L(v)} B^-_v(w)$ to be revealed before f_s.

Consequently the events A_E partition A. We furthermore define the set $A^-_{v,E}(w)$, the function $g_{v,w,E}(vw)$ and the event $F(w, E)$ as follows. We set $A^-_{v,E}(w)$ to be a subset of $A^-_v(w)\backslash E$ of size $\frac{\log n}{100} - r$ and we define the map $g_{v,w,E} : A^-_{v,E}(w) \to \left[\frac{\log n}{100} - r\right]$ given by the relation $g_{v,w,E}(xw) = k$ where xw is the kth arc that was revealed out of the arcs in $A^-_{v,E}(w)$. In addition we set $F(w, E)$ to be the event that A_E occurs and that in $D^{(1)}_\tau$ $\nexists \ell \in \mathbb{Z}_{\geq 0}$ s.t. $\ell q + q < g_{v,w,E}(vw)$ and $g^{-1}_{v,w,E}(\ell q + 1), ..., g^{-1}_{v,w,E}(\ell q + q)$ are colored by q distinct colors.
For $E \in \mathcal{E}_1$, suppose we condition on A_E. By using the same tools as in Lemma 16 with $A^-_{v,E}(\cdot)$, $g_{v,\cdot,E}(v\cdot)$ and $F(\cdot, E)$ in place of $A_v(\cdot)$, $g_{v,\cdot}(v\cdot)$ and $F(\cdot)$ respectively, we have that for $w \in N^+_L(v)\backslash\{v^*\}$ where $vv^* = f_E$ the events $F(w, E)$ occur independently with probability at most $\frac{q^{q+1}}{\frac{\log n}{100} - r}$. On the other hand f_E contributes to the the coloring of v with probability 1. Therefore, as the events A_E partition A, the probability that fewer than q arcs contribute to the coloring of v in $D^{(1)}_\tau$ *conditioned on the event* A is bounded above by

$$\mathbb{P}\left(Bin\left(d^+(v) - 1, \frac{q^{q+1}}{\frac{\log n}{100} - r}\right) \geq [d^+(v) - 1] - (q - 2)\right).$$

As $\mathbb{E}\left[|A^+_L(v) \cap \{h_v^{-1}(1), h_v^{-1}(2), ..., h_v^{-1}(r)\}|\right] = \frac{r}{|A^+_L(v)|\frac{\log n}{100}} \cdot |A^+_L(v)|$, Markov's inequality gives

$$\mathbb{P}(A) = \mathbb{P}\big(|A^+_L(v) \cap \{h_v^{-1}(1), h_v^{-1}(2), ..., h_v^{-1}(r)\}| \geq 1\big) \leq \frac{100r}{\log n}.$$

• Case 2: The event $A^c \cap B$ occurs.
Set $\mathcal{E}_2 = \left\{(f_1, ...f_r) \in \left(\bigcup_{w\in A^+_L(v)} B^-_v(w)\right)^r : f_1, ..., f_r \text{ are distinct and } |\{f_1, ..., f_r\} \cap A^-_v(w)| <\right.$

r_ℓ for some $w \in N_L^+(v)$$\Big\}$. Henceforth we can proceed as in Case 1 but without using the guaranteed contribution of the first arc in $A_L^+(v)$. Thus, conditioned on the event $A^c \cap B$, the probability that fewer than q arcs contribute to the coloring of v in $D_\tau^{(1)}$ is bounded above by

$$\mathbb{P}\bigg(Bin\bigg(d^+(v), \frac{q^{q+1}}{\frac{\log n}{100} - r}\bigg) \geq d^+(v) - (q-1)\bigg).$$

Furthermore,

$$\begin{aligned}
\mathbb{P}(A^c \cap B) \leq \mathbb{P}(B) &\leq d^+(v) \sum_{i=0}^{r_\ell - 1} \binom{\frac{\log n}{100}}{i} \binom{(d^+(v)-1)\frac{\log n}{100}}{r-i} \Big/ \binom{d^+(v)\frac{\log n}{100}}{r} \\
&= d^+(v) \sum_{i=0}^{r_\ell - 1} \binom{\frac{\log n}{100}}{i} \binom{(d^+(v)-1)\frac{\log n}{100}}{r-i} \binom{r}{r-i} \Big/ \binom{d^+(v)\frac{\log n}{100}}{r-i} \binom{d^+(v)\frac{\log n}{100} - r + i}{i} \\
&\leq d^+(v) \sum_{i=0}^{r_\ell - 1} r^i \binom{(d^+(v)-1)\frac{\log n}{100}}{r-i} \Big/ \binom{d^+(v)\frac{\log n}{100}}{r-i} \\
&\leq d^+(v) \sum_{i=0}^{r_\ell - 1} r^i \prod_{j=0}^{r-i-1} \frac{(d^+(v)-1)\frac{\log n}{100} - j}{d^+(v)\frac{\log n}{100} - j} \leq d^+(v) \sum_{i=0}^{r_\ell - 1} r^i \bigg(\frac{(d^+(v)-1)\frac{\log n}{100}}{d^+(v)\frac{\log n}{100}}\bigg)^{r-r_l} \\
&\leq d^+(v) \cdot r^{r_\ell} \cdot \exp\bigg\{ - \frac{r - r_l}{d^+(v)} \bigg\} \\
&\leq \exp\bigg\{ \log\big(d^+(v)\big) + q^q \log\log n \cdot 5\log(\log\log n) - 0.4(\log\log n)^4 \bigg\} = o\bigg(\frac{1}{\log^3 n}\bigg).
\end{aligned}$$

To get from the second to the third line we are using the fact that $d^+(v) \geq 2$. Furthermore at the last inequality we use that $d^+(v) \leq \log\log n$.

- Case 3: The event $A^c \cap B^c$ occurs.

Set $\mathcal{E}_3 = \Bigg\{ (f_1, ...f_r) \in \bigg(\bigcup_{w \in A_L^+(v)} B_v^-(w) \bigg)^r : f_1, ..., f_r$ are distinct and for every $w \in N_L^+(v)$ we have that $|\{f_1, ..., f_r\} \cap B_v^-(w)| \geq r_\ell \Bigg\}$. For $E \in \mathcal{E}_3$ we let A_E be the event that for all $i \in [r]$, f_i is the i-th edge that is revealed from $\bigcup_{w \in A_L^+(v)} A_v^-(w)$. Consequently we have that the events A_E partition the event $A^c \cap B^c$. Furthermore for $E = (f_1, ...f_r) \in \mathcal{E}_3$ and $w \in N_L^+(v)$ we set $\tilde{A}_{v,E}^-(w)$ to be a subset of $A_v^-(w)$ of size $\frac{\log n}{100} - r + r_\ell$ such that $|\tilde{A}_{v,E}^-(w) \cap \{e_1, ..., e_r\}| = r_\ell$ and define the map $\tilde{g}_{v,w,E} : \tilde{A}_{v,E}^-(w) \mapsto \big[\frac{\log n}{100} - r + r_\ell\big]$ and the event $\tilde{F}(w, E)$ correspondingly. Note that for $w \in N_L^+(v)$ and for $E \in \mathcal{E}_3$ since $A_E \subseteq A^c \cap B^c$ we have that $\tilde{g}_{v,w,E}(vw) > r_\ell$. Thus, as in the proof of Lemma 16 for any $E \in \mathcal{E}_3$ and $w \in N_L^+(v)$ we have,

$$\mathbb{P}\big(\tilde{F}(w,E) | A_E\big) = \sum_{k=r_\ell+1}^{\frac{\log n}{100} - r + r_\ell} \mathbb{P}\big(\tilde{g}_{v,w,E}(vw) = k \wedge \tilde{F}(w,E) | A_E\big)$$

$$= \sum_{k=r_\ell+1}^{\frac{\log n}{100}-r+r_\ell} \mathbb{P}\big(\tilde{g}_{v,w,E}(vw)=k|A_E\big)\mathbb{P}\bigg(\tilde{F}(w,E)|\tilde{g}_{v,w}(vw)=k\wedge A_E\bigg)$$
$$\leq \sum_{k=r_\ell}^{\frac{\log n}{100}} \frac{1}{\frac{\log n}{100}-r}\bigg(1-\frac{1}{q^q}\bigg)^{\lfloor k/q\rfloor} \leq \sum_{j\in\mathbb{N}} \frac{100}{\log n-100r}\bigg(1-\frac{1}{q^q}\bigg)^{\lfloor r_\ell\rfloor}\bigg(1-\frac{1}{q^q}\bigg)^{j}$$
$$\leq \sum_{j\in\mathbb{N}} \frac{101}{\log n}\cdot\exp\bigg(-\frac{1}{q^q}\cdot\lfloor q^q\log\log n\rfloor\bigg)\cdot\bigg(1-\frac{1}{q^q}\bigg)^{j} \leq \frac{101eq^q}{\log^2 n}.$$

Once more, for fixed $E\in\mathcal{E}_3$, conditioned on A_E the events $F(w,E)$ are independent (as in case 1). Furthermore the events A_E for $E\in\mathcal{E}_3$ partition $A^c\cap B^c$. Hence, conditioned on the occurrence of event $A^c\cap B^c$ the probability that less than q arcs contribute to the coloring of v in $D_\tau^{(1)}$ is bounded by

$$P\bigg(Bin\bigg(d^+(v),\frac{101eq^q}{\log^2 n}\bigg)\geq d^+(v)-(q-1)\bigg).$$

Finally, by conditioning on the occurrence of event A or $A^c\cap B$ or $A^c\cap B^c$ we get that for a vertex v in $D_\tau^{(1)}$ satisfying $q\leq d^+(v)\leq\log\log n$ we have,

$$\begin{aligned}
&\mathbb{P}\big(\text{fewer than } q \text{ arcs contribute to the coloring of } v \text{ in } D_\tau^{(1)}\big)\\
&\leq \mathbb{P}\bigg(Bin\bigg(d^+(v)-1,\frac{100q^{q+1}}{\log n-100(\log\log n)^5}\bigg)\geq [d^+(v)-1]-(q-2)\bigg)\frac{100(\log\log n)^5}{\log n}\\
&+\mathbb{P}\bigg(Bin\bigg(d^+(v),\frac{100q^{q+1}}{\log n-100(\log\log n)^5}\bigg)\geq d^+(v)-q+1\bigg)\frac{1}{\log^3 n}\\
&+\mathbb{P}\bigg(Bin\bigg(d^+(v),\frac{101eq^q}{\log^2 n}\bigg)\geq d^+(v)-q+1\bigg)\\
&\leq \frac{101(\log\log n)^5}{\log n}\binom{d^+(v)-1}{q-2}\bigg(\frac{101q^{q+1}}{\log n}\bigg)^{d^+(v)-q+1}. \quad \square
\end{aligned}$$

Lemma 18. *Let $e_{t_q}=(v^*,w^*)$ be such that $d^+(v^*)=q$. Then probability that fewer than q arcs contribute to the coloring of v^* in $D_\tau^{(1)}$ is bounded above by*

$$\frac{101(\log\log n)^5}{\log n}\binom{d^+(v)-1}{q-2}\bigg(\frac{101q^{q+1}}{\log n}\bigg)^{d^+(v)-q+1}.$$

Proof. As seen in the proof of Lemma 16 every arc out of v^* except (v^*w^*) contributes to the coloring of v with probability $=\frac{100q^{q+1}}{\log n}$. Thereafter since $g_{v,w}(v^*w^*)=\frac{\log n}{100}$ the first line of (3.1) gives as

$$\mathbb{P}(F(w^*))\leq \prod_{l=1}^{\lfloor\frac{\log n}{100}/q\rfloor}\bigg(1-\frac{1}{q^q}\bigg)\leq\bigg(1-\frac{1}{q^q}\bigg)^{\frac{\log n}{100q}}\leq e^{-\frac{\log n}{100q^{q+1}}}\leq\frac{100q^{q+1}}{\log n}.$$

Therefore the probability that fewer than q arcs contribute to the coloring of v^* in $D_\tau^{(1)}$ is bounded above by $q \cdot \frac{100q^{q+1}}{\log n} \leq \frac{101(\log\log n)^5}{\log n}\binom{d^+(v)-1}{q-2}\left(\frac{101q^{q+1}}{\log n}\right)^{d^+(v)-q+1}$.

□

Proof of Theorem 2: We say *DCOL* fails if once the last edge has been revealed, there exist a vertex $v \in V$ and a color $c \in [q]$ such that the in- or out-degree of v in color c is 0. Observed that conditioned on the almost sure event $\{m_\ell \leq \tau\}$ Lemma 11 implies that for all $k \in [q, 3\log n \backslash \log\log n]$ the number of vertices of degree at most k is at most $v_k = e^{2\omega(n)}(\log n)^{k-q+1}/(k-1)!$. Thus from Lemmas 15, 16, 17 and Remark 3.4.2, by implicitly conditioning on the event $\{m_\ell \leq \tau\}$ and Lemma 11, we have

$$\begin{aligned}
\mathbb{P}(\text{DCOL fails}) &\leq 2\mathbb{P}\big(\exists v \in D_{m_\ell} \text{ such that } < q \text{ arcs contribute to the coloring of } v \text{ in } D_{m_\ell}^{(1)}\big)\\
&\leq 2\sum_{k=\frac{3\log n}{\log\log n}}^{n} n\cdot\binom{k-1}{q-1}\left(\frac{100q^{q+1}}{\log n}\right)^{k-q} + 2\sum_{k=\log\log n+1}^{\frac{3\log n}{\log\log n}} v_k\cdot\binom{k-1}{q-1}\left(\frac{100q^{q+1}}{\log n}\right)^{k-q}\\
&\qquad + 2\sum_{k=q}^{\log\log n} v_k\cdot\frac{101(\log\log n)^5}{\log n}\binom{k-1}{q-2}\left(\frac{101q^{q+1}}{\log n}\right)^{k-q+1}\\
&\leq 2\sum_{k=\frac{3\log n}{\log\log n}}^{n} n\cdot k^q\left(\frac{100q^{q+1}}{\log n}\right)^{k-q} + 2\sum_{k=\log\log n}^{\frac{3\log n}{\log\log n}} \frac{e^{2\omega}(\log n)^{k-q+1}}{(k-1)!}\cdot\binom{k-1}{q-1}\left(\frac{100q^{q+1}}{\log n}\right)^{k-q}\\
&\qquad + 2\sum_{k=q}^{\log\log n} \frac{e^{2\omega(n)}(\log n)^{k-q+1}}{(k-1)!}\cdot\frac{101(\log\log n)^5}{\log n}q\binom{k-1}{q-1}\left(\frac{101q^{q+1}}{\log n}\right)^{k-q+1}\\
&\leq 2\sum_{k=\frac{3\log n}{\log\log n}}^{n}\frac{1}{n^2} + 2\sum_{k=\log\log n}^{\frac{3\log n}{\log\log n}}\frac{e^{2\omega(n)}\log n}{(k-q)!}(100q^{q+1})^{k-q}\\
&\quad + 2\frac{101^2q^{q+2}(\log\log n)^5\cdot e^{2\omega(n)}}{\log n}\left[\sum_{k=q+1+202eq^{q+1}}^{\log\log n}\frac{(101q^{q+1})^{k-q}}{(k-q)!} + \sum_{k=q}^{q+202eq^{q+1}}\frac{(101q^{q+1})^{k-q}}{(k-q)!}\right]\\
&\leq \frac{2}{n} + 2\log^2 n\sum_{k=\log\log n}^{\frac{3\log n}{\log\log n}}\left(\frac{100q^{q+1}e}{k-q}\right)^{k-q}\\
&\qquad + \frac{C_1(\log\log n)^6}{\log n}\left[\sum_{k=q+1+202eq^{q+1}}^{\log\log n}\left(\frac{101q^{q+1}e}{(k-q)}\right)^{k-q} + C_2\right]\\
&\leq \frac{2}{n} + 2\log^2 n\cdot\frac{3\log n}{\log\log n}\left(\frac{100q^{q+1}e}{\log\log n - q}\right)^{\log\log n-q} + O\left(\frac{(\log\log n)^6}{\log n}\right) = o(1),
\end{aligned}$$

for some sufficiently large constants $C_1 = C_1(q)$ and $C_2 = C_2(q)$ depending only on q. □

3.5 Finding Hamilton cycles - Overview

We may now proceed to show that w.h.p. for every color $c \in [q]$, $DCOL$ succeeds in assigning color c to every edge in some Hamilton cycle in D_τ. We set D'_c to be the subgraph of D_τ induced by the edges of color c. We start by constructing a minor D_c of D'_c. To do so we first remove some arcs and then applying contractions to arcs adjacent to vertices in BAD. By doing the contractions we hide the vertices in BAD while the arc removal ensures that any Hamilton cycle in D_c also yields a Hamilton cycle in D'_c.

We organize the rest of the proof as follows. We first deal with *Phase 1* which takes place in our original setting. We then give a reduction of Theorem 1 to Lemma 10. Finally we explicitly describe *Phases 2 & 3* and use them to prove Lemma 10. *Phases 2 & 3* take place in the more general setting of Lemma 10.

During *Phase 1* we use out-arcs and in-arcs that have been revealed during the time intervals $(m_1, m_2]$ and $(m_2, m_3]$ respectively in order to show that w.h.p. there exists a matching in D_c consisting of at most $2 \log n$ cycles spanned by E_{m_3}. By matching we refer to a complete matching i.e. some $M \subseteq V_c \times V_c \backslash \{(v, v) : v \in V_c\}$ where every vertex has in- and out-degree exactly 1.

Thereafter, we randomly partition $E' = E^2 \cup E^3$. In *Phase 2*, we attempt to sequentially join any two cycles found in the current matching, starting with the matching above, to a single one. We join the cycles by a straightforward two-arc exchange, where arcs vw, xy in two distinct cycles are rerouted via vy, xw if the latter two are in E^2 (illustrated at Figure 2). We show that once this is no longer possible, we are left with a large cycle consisting of $n - o(n)$ vertices of D_τ.

Finally, during *Phase 3* using arcs found in E^3, we sequentially try to merge the smaller cycles with the largest one. To merge two cycles here we start by finding an arc in E^3 joining them. This creates a dipath spanning the vertices of the two cycles. Afterwards, we grow the set of dipaths using "double rotations", or sequences of two-arc exchanges that maintain a dipath on the same vertex set. (More specifically, for a dipath $P = (p_1, p_2, ..., p_s)$, suppose $p_s p_k, p_{k-1} p_l \in E^3$ with $k < l$. Then a double rotation, illustrated at Figure 2, using those two arcs replaces P with the dipath $P' = (p_1, p_2, ..., p_{k-1}, p_l, p_{l+1}, ..., p_s, p_k, p_{k+1}, ..., p_{l-1})$.) By performing sequences of double rotations we find $\Omega(n)$ paths with a common starting vertex but *distinct* endpoints. With this many paths we succeed in closing one of them (joining the end-vertex to the start-vertex by an arc) with probability at least $1 - o(n^{-\epsilon})$ for some $\epsilon > 0$. Hence we may join all ($\leq 2 \log n$) cycles inherited from *Phase 2*.

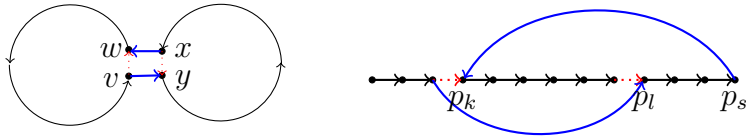


Figure 3.2: Left-Merging two cycles (*Phase 2*), Right-Double rotation (*Phase 3*).

3.6 Construction of D_c

Let D'_c be the graph induced by the arcs of color c, $BAD = \{z_1, z_2, ..., z_b\}$ where for some $s \leq b$ we have that $SMALL \cap BAD = \{z_1, z_2, ...z_s\}$. D_c is set to be the graph that we get after applying the following algorithm to D'_c. We aim to thread all BAD vertices, one at a time, into disjoint directed paths (we will later contract) with neither endpoint in BAD. We achieve this by dynamically keeping track of all potential starting vertices V^+ and potential ending vertices V^- of these paths. It is likely that some BAD vertices will have been used as endpoints of paths for other BAD vertices before they had their turn-see the "if" clause below-but, in this case, we only need to extend the path in a single direction.

Algorithm 4 HideBad

$V^+ := V_n$, $V^- := V_n$, $E_{contr} := \emptyset$.
for $\ell = 1, 2, ..., s$ **do**
  Let $j, k \in [n]$ be minimal such that $v_j \in V^+, v_k \in V^-$ and $v_j z_\ell, z_\ell v_k \in E(D'_c)$
  $V^+ \leftarrow V^+ \backslash \{z_\ell, v_j\}, V^- \leftarrow V^- \backslash \{z_\ell, v_k\}, E_{contr} \leftarrow E_{contr} \cup \{v_j z_\ell, z_\ell v_k\}$.
end
for $\ell = s+1, s+2, ..., b$ **do**
  if $z_\ell \notin V^+$ **then**
    Let $j \in [n]$ be the minimum such that $v_j \in V^+$ and $v_j z_\ell \in E(D'_c)$
    $V^+ \leftarrow V^+ \backslash \{v_j\}, V^- \leftarrow V^- \backslash \{z_\ell\}, E_{contr} \leftarrow E_{contr} \cup \{v_j z_\ell\}$.
  else if $z_\ell \notin V^-$ **then**
    Let $k \in [n]$ be the minimum such that $v_k \in V^-$ and $z_\ell v_k \in E(D'_c)$
    $V^+ \leftarrow V^+ \backslash \{z_\ell\}, V^- \leftarrow V^- \backslash \{v_k\}, E_{contr} \leftarrow E_{contr} \cup \{z_\ell v_k\}$.
  else
    Let $j, k \in [n]$ be minimal such that $v_j \in V^+, v_k \in V^-$ and $v_j z_\ell, z_\ell v_k \in E(D'_c)$
    $V^+ \leftarrow V^+ \backslash \{z_\ell, v_j\}, V^- \leftarrow V^- \backslash \{z_\ell, v_k\}, E_{contr} \leftarrow E_{contr} \cup \{v_j z_l, z_\ell v_k\}$.
  end
end
Delete all arcs xy in $E(D'_c) \backslash E_{contr}$ such that $x \notin V^+$ or $y \notin V^-$.
Contract all edges in E_{contr} and let D_c be the resultant graph.

It should not be obvious at this stage that we can always perform this algorithm so greedily, as one could feasibly run out of potential out- or in-neighbours of a given $z_\ell \in BAD$ at some late stage, all taken up by earlier BAD vertices. We will devote the rest of this section to showing (Theorem 3) this is unlikely to be a problem (after reassuring ourselves that Hamiltonicity is preserved under these contractions in Lemma 20).

Remark 3.6.1. At each step of the algorithm $x \in V_n$ is removed from V^+ (similarly from V^-) iff for some $y \in V_n$ the arc xy (yx respectively) is added to E_{contr}.

Notation 3.6.2. Henceforth we denote by V_c the vertex set of D_c.

Definition 3.6.3. For $v \in V_c$ set $contr(v)$:=$\{u \in V(D'_c)$: u gets contracted to $v\}$. Furthermore set v^+ and v^- to be the unique elements found in $contr(v) \cap V^+$ and $contr(v) \cap V^-$ respectively.

Remark 3.6.4. Every $v \in V_c$ has both $v^+, v^- \notin BAD$. Furthermore $V^* := V \setminus (BAD \cup N(BAD)) \subseteq V^+, V^-$.

Lemma 19. *For $u, v \in V_c$ we have that $uv \in E(D_c) \Leftrightarrow u^+v^- \in E(D'_c)$.*

Proof. Observe that $xy \in E(D'_c)$ was removed or contracted iff after the last iteration of *HideBad* $x \notin V^+$ or $y \notin V^-$. Let $u, v \in V_c$ be such that $u^+v^- \in E(D'_c)$. Then since $u^+ \in V^+$ and $u^- \in V^-$, from the observation follows that u^+v^- was not removed or contracted. In addition u^+v^- is identified with uv after the contractions, hence $uv \in E(D_c)$. Let $a, b \in V_c$ be such that $ab \in E(D_c)$ so certainly $a \neq b$. ab originated from an edge in $\big(contr(a) \times contr(b)\big) \cap E(D'_c)$ and since any edge in $\big(contr(a) \times contr(b)\big) \setminus \{a^+b^-\}$ was either contracted or removed it must be the case that $u^+v^- \in E(D'_c)$. □

Lemma 20. *If there exists a Hamilton cycle in D_c then there exists a Hamilton cycle in D'_c.*

Proof. For $u \in V_c$ define $P(u)$ to be the dipath in D'_c that contains all the vertices in $contr(u)$, starts at u^-, ends with u^+ and uses all the arcs in E_{contr} that are spanned by $contr(u)$ (in the case that $|contr(u)| = 1$, $P(u)$ is a single vertex i.e. a dipath of length 0). Now suppose $v_{\pi(1)}, v_{\pi(1)}v_{\pi(2)}, v_{\pi(2)}, ..., v_{\pi(n_c)}, v_{\pi(n_c)}v_{\pi(1)}, v_{\pi(1)}$ is a Hamilton cycle in D_c then, we have that $P(v_{\pi(1)}), v^+_{\pi(1)}v^-_{\pi(2)}, P(v_{\pi(2)}), ..., P(v_{\pi(n_c)}), v^+_{\pi(n_c)}v^-_{\pi(1)}, P(v^-_{\pi(1)})$ is a Hamilton cycle in D'_c. To see this, first note that $P(v_{\pi(i)})$ starts with $v^-_{\pi(i)}$ and ends with $v^+_{\pi(i)}$. Moreover $v_{\pi(i)}v_{\pi(i+1)} \in E(D_c)$ implies, by Lemma 19, that $v^+_{\pi(i)}v^-_{\pi(i+1)} \in E(D'_c)$. Finally, since the sets $contr(v)$ partition V_n, each vertex in V_n appears exactly in one of the dipaths $P(u)$. □

Theorem 3. *W.h.p. the algorithm HideBad terminates.*

The proof of Theorem 3 will follow from Lemmas 21 and 23 proven in this section. To state and prove these we will need the following definitions.

Definition 3.6.5. For $v \in V_n$, let $N(v)$:=$\{u \in V_n : d'_\tau(u, v) = 1\}$ (i.e those vertices whose undirected distance from v is one). Similarly set $N(N(v))$:=$\{u \in V_n : d'_\tau(u, v) \in \{1, 2\}\}$.

Remark 3.6.6. All three sets of edges that appear at times found in $(0, m_1], (m_1, m_2]$ and $(m_2, m_3]$ respectively are distributed as the edges of D_{n,m_1}. Hence, by additionally taking into account the symmetry between in- and out- arcs in D_{n,m_1}, the sets B_1^+, B_1^-, B_2^+ and B_3^- (defined during the execution of *DCOL*) follow the same distribution.

Lemma 21. *W.h.p. for all $v \in V_n$ we have that $|BAD \cap N(N(v))| \leq 4e^{q \cdot 10^5}$.*

Proof. Let $k = e^{q \cdot 10^5}$ and suppose $|BAD \cap N(N(v))| > 4k$ for some $v \in V_n$. Then there is some digraph $S \subseteq D_\tau$ with $V(S) = \{v, b_1, ..., b_k, w_1, ..., w_l\}$ for some $l \leq k$ satisfying the

following . For some $i \leq k$ all of the vertices $b_1, ..., b_i, w_1, ..., w_l$ are connected to v by arcs $e_1, ..., e_{i+l}$ and for $i < j \leq k$, b_j is connected to some $v_j \in \{b_1, ..., b_i, w_1, ..., w_l\}$ by the arc e_{j+l}. Furthermore there is some $B^* \in \{B_1^+, B_1^-, B_2^+, B_3^-\}$ such that $B = \{b_1, ..., b_k\} \subseteq B^*$. Suppose $B^* = B_1^+$. By setting for $E \subseteq E(S)$ the events $S_{m_1}(E) := \{E(S) \cap E_{m_1} = E\}$ and $S_{m_1,\tau}(E) := \{E(S) \backslash E \subseteq E_\tau \backslash E_{m_1}\}$ we have,

$$L = \mathbb{P}\big(\{S \subseteq D_\tau\} \wedge \{B \subseteq B_1^+\}\big) = \sum_{E \subseteq E(S)} \mathbb{P}\big(S_{m_1}(E) \wedge S_{m_1,\tau}(E) \wedge \{B \subseteq B_1^+\}\big)$$
$$= \sum_{E \subseteq E(S)} \mathbb{P}\big(S_{m_1}(E)\big) \cdot \mathbb{P}\big(S_{m_1,\tau}(E) \big| S_{m_1}(E)\big) \cdot \mathbb{P}\big(B \subseteq B_1^+ \big| S_{m_1}(E) \wedge (S_{m_1,\tau}(E)\big).$$

For fixed $E \subseteq E(S)$ (1) implies that,

$$\mathbb{P}\big(S_{m_1}(E)\big) \leq 10\sqrt{m_1} p_1^{|E|}(1-p_1)^{|E(S)\backslash E|} \leq n p_1^{|E|} \leq n\left(\frac{\log n}{n}\right)^{|E|}.$$

Furthemore,

$$\mathbb{P}\big(S_{m_1,\tau}(E)\big|S_{m_1}(E)\big) = \frac{\binom{n(n-1)-m_1-|E(S)\backslash E|}{\tau-m_1-|E(S)\backslash E|}}{\binom{n(n-1)-m_1}{\tau-m_1}} = \frac{\binom{\tau-m_1}{|E(S)\backslash E|}}{\binom{n(n-1)-m_1}{|E(S)\backslash E|}} = \prod_{i=0}^{|E(S)\backslash E|-1} \frac{\tau-m_1-i}{n(n-1)-m_1-i}$$
$$\leq \left(\frac{\tau-m_1}{n(n-1)-m_1}\right)^{|E(S)\backslash E|} \leq \left(\frac{2n\log n}{n^2}\right)^{|E(S)\backslash E|}.$$

Finally, in order to bound $\mathbb{P}\big(B \subseteq B_1^+ \big| S_{m_1}(E) \wedge (S_{m_1,\tau}(E)\big)$ from above note the following. There are $\binom{n(n-1)-|E(S)|}{m_1-|E|}$ ways to pick $E_{m_1} \backslash E$ so that it can be extended to a chain $E_{m_1} \backslash E \subseteq E_{m_1} \backslash E_\tau$ such that E_{m_1} and E_τ satisfy both the events $S_{m_1}(E)$ and $S_{m_1,\tau}(E)$. Given $S_{m_1}(E)$ and $S_{m_1,\tau}(E)$ occur $E_{m_1} \backslash E$ is equally likely to be any of those $\binom{n(n-1)-|E(S)|}{m_1-|E|}$ choices. Moreover, if $B \subseteq B_1^+$ then every vertex in B has at most $\epsilon \log n$ out-arcs in E_{m_1}. Hence there are at most $f = \epsilon|B|\log n = \epsilon k \log n$ arcs in $E_{m_1} \backslash E$ with out-vertex in B (i.e. from the set $\{bv : b \in B, v \in V_n \text{ and } v \neq b\}$). Thus,

$$\mathbb{P}\big(B \subseteq B_1^+ \big| S_{m_1}(E) \wedge S_{m_1,\tau}(E)\big) \leq \frac{\sum_{j=0}^{f} \binom{k(n-1)}{j}\binom{n(n-1)-k(n-1)}{m_1-|E|-j}}{\binom{n(n-1)-|E(S)|}{m_1-|E|}} \leq \frac{f\binom{k(n-1)}{f}\binom{n(n-1)-k(n-1)}{m_1-|E|-f}}{\binom{n(n-1)-|E(S)|}{m_1-|E|}}$$
$$\leq f\binom{k(n-1)}{f} \frac{(m_1-|E|)!}{(m_1-|E|-f)!} \frac{\frac{[n(n-1)-k(n-1)]!}{[n(n-1)-k(n-1)-m_1+|E|+f]!}}{\frac{\left(\prod_{j=0}^{f-1} n(n-1)-|E(S)|-m_1+|E|+f-j\right)[n(n-1)-|E(S)|]!}{[n(n-1)-|E(S)|-m_1+|E|+f]!}}$$
$$\leq f\left(\frac{ekn}{f}\right)^f \prod_{j=0}^{f-1} \frac{m_1-|E|-j}{n(n-1)-|E(S)|-m_1+|E|+f-j} \prod_{j=0}^{m_1-|E|-f-1} \frac{n(n-1)-k(n-1)-j}{n(n-1)-|E(S)|-j}$$
$$\leq f\left(\frac{ekn}{f}\right)^f \left(\frac{m_1}{0.9n^2}\right)^f \cdot \prod_{j=0}^{m_1-|E|-f-1} \frac{n(n-1)-k(n-1)}{n(n-1)-|E(S)|}$$

$$\leq f\left(\frac{ekm_1}{0.9fn}\right)^f \exp\left\{-\frac{k(n-1)-|E(S)|}{n(n-1)}\cdot(m_1-|E(S)|-f-1)\right\}$$
$$\leq \epsilon k\log n\left(\frac{ekm_1}{0.9\epsilon kn\log n}\right)^{\epsilon k\log n}\exp\left\{-\frac{0.8km_1}{n}\right\}\leq \epsilon k\log n\left(\frac{1}{\epsilon}\right)^{\frac{m_1}{n}}\exp\left\{-\frac{0.8km_1}{n}\right\}$$
$$\leq \epsilon k\log n\cdot\exp\left\{[-\log(\epsilon)-0.8k]\frac{m_1}{n}\right\}\leq \exp\left\{-0.7k\cdot\frac{m_1}{n}\right\}$$
$$\leq \exp\left\{-0.7e^{q\cdot 10^5}e^{-q\cdot 10^4}\log n\right\}\leq \exp\left\{-e^{8.9q\cdot 10^4}\log n\right\}.$$

The 2nd inequality follows from the fact that $\binom{k(n-1)}{j}\binom{n(n-1)-k(n-1)}{m_1-|E|-j}$ is increasing for $j\in[1,f]$. Thus, using the upper bounds found for the quantities on the right hand side of (3.6) we obtain

$$L\leq\sum_{E\subseteq E(S)} n\left(\frac{\log n}{n}\right)^{|E|}\cdot\left(\frac{2\log n}{n}\right)^{|E(S)\setminus E|}\cdot\exp\left\{-e^{8.9q\cdot 10^4}\log n\right\}$$
$$\leq\sum_{E\subseteq E(S)}\left(\frac{\log n}{n}\right)^{|E(S)|}\exp\left\{-e^{8.8q\cdot 10^4}\log n\right\}\leq\left(\frac{\log n}{n}\right)^{|E(S)|}\exp\left\{-e^{8q\cdot 10^4}\log n\right\}.$$

For fixed l,k there are exactly $n\binom{n-1}{k}\binom{n-1-k}{l}$ ways to choose the vertices of S, or equivalently, disjoint sets $\{v\},\{b_1,...,b_k\}$ and $\{w_1,...,w_l\}$ from V_n. Thereafter there are at most $2^{l+k}\sum_{i=0}^k\binom{k}{i}(i+l)^{k-i}$ choices for its directed edges. Taking into account Remark 3.6.6 and that $l\leq k=e^{q\cdot 10^4}$, union bound gives us

$$\mathbb{P}\big(\exists v\in V_n:|BAD\cap N(N(v))|>4e^{q\cdot 10^5}\big)$$
$$\leq\mathbb{P}\big(\exists v\in V_n\text{ and }(i,*)\in\{(1,+),(1,-),(2,+),(3,-)\}:B_i^*\cap N(N(v))>e^{q\cdot 10^5}\big)$$
$$\leq 4\sum_{l=0}^k n\binom{n-1}{k}\binom{n-1-k}{l}2^{k+l}\sum_{i=0}^k\binom{k}{i}(i+l)^{k-i}\left(\frac{\log n}{n}\right)^{l+k}\exp\left\{-e^{8q\cdot 10^4}\log n\right\}$$
$$\leq 4\sum_{l=0}^k n^{l+k+1}\left(\frac{\log n}{n}\right)^{l+k}\exp\left\{-e^{7q\cdot 10^4}\log n\right\}=o(n^{-2}).\qquad\square$$

Lemma 22. *W.h.p. for every $u\notin BAD$ we have that $u\in FULL^+_{m_1}\cap FULL^-_{m_1}$.*

Proof. With $k=4e^{q\cdot 10^5}$ Lemma 21 implies that w.h.p. for every $u\in V_n$ we have $|\{w\in V_n:$ $w\in N^+(u)$, $h(uw)<m_1$ and $d^+_{m_1}(w)\leq\epsilon\log n\}|\leq k$. Hence as $u\notin BAD$ implies that $d=\min\{d^+_{m_1}(u),d^-_{m_1}(u)\}\geq\epsilon\log n$ from Remark 3.4.11, with $\gamma=\epsilon$ it follows that

$$\mathbb{P}\big(\exists u\notin BAD:u\notin FULL^+_{m_1}\cap FULL^-_{m_1}\big)\leq 2n\max_{\epsilon\log n\leq d\leq n}\left\{\binom{d-k}{q-1}\left(\frac{q^{q+1}}{\epsilon\log n}\right)^{(d-k)-(q-1)}\right\}$$
$$\leq 2nn^{q-1}\left(\frac{q^{q+1}}{\epsilon\log n}\right)^{0.5\epsilon\log n}=o(1).\qquad\square$$

Lemma 23. *W.h.p. for every $v \in BAD\backslash SMALL$ we have that v has at least $\log\log n$ out-arcs in each color ending in $V_n\backslash BAD$ and at least $\log\log n$ in-arcs in each color starting from $V_n\backslash BAD$.*

Proof. Let $v \in BAD\backslash SMALL$. Then v has at least $\frac{\log n}{100}$ out-neighbors. Lemma 14 gives us that the out- degree of v at time m_3 is at most $\frac{3\log n}{10^3 q}$. Therefore v has at least $\frac{\log n}{100} - \frac{3\log n}{10^3 q} - 4e^{q\cdot 10^5}$ out-neighbors in $V_n\backslash BAD$ that arrive after m_3. By the previous Lemma w.h.p. for all $u \in V_n\backslash BAD$ and all $c \in [q]$ we have $d^-_{m_1}(u,c) \geq 1$. Hence at most q such arcs vu that arrive at some time $t > m_3$ will be colored under the condition $v \notin FULL^+_{t-1}$. Thus there are at least $\frac{\log n}{100} - \frac{3\log n}{10^3 q} - 4e^{q\cdot 10^5} - q$ arcs vu with $u \in V_n\backslash BAD$ that will arrive at some time $t > m_3$ and will be colored with color c that minimizes $d^+_t(v,c)\mathbb{I}\{v \in BAD\} + d^-_t(u,c)\mathbb{I}\{u \in BAD\} = d^+_t(v,c)$ (i.e. the arcs are given a color in which v has the smallest out-degree when they appear). Thus v will have at least $\frac{1}{q}\big(\frac{\log n}{100} - \frac{3\log n}{10^3 q} - 4e^{q\cdot 10^5} - q\big) - 1 \geq \log\log n$ out-arcs in each color ending in $V_n\backslash BAD$. A similar argument holds for the number of arcs from $V_n\backslash BAD$ to v. □

Proof of Theorem 3. Assume that the algorithm *HideBad* does not terminate. Then there is an iteration f at which there do not exist $v_j \in V^+$ and $v_k \in V^-$ such that $v_j z_f, z_f v_k \in E(D'_c)$, WLOG the former (the case $\nexists v_k \in V^-$ will follow similarly).

Case 1: $f \leq s$ (i.e $z_f \in SMALL$). As every vertex has in-degree at least one in D'_c, there exists $x \in V_n$ such that the arc xz_f belongs to E_τ and has color c. Hence, $\exists \ell < f$ such that at ℓ-th iteration x was removed from V^+. This implies that $z_\ell \in N(N(z_f))$. Hence we get that z_ℓ, z_f belong to $SMALL$ and z_f, z_ℓ have distance less than 3 contradicting Lemma 12.

Case 2: $s < f \leq b$ (i.e $z_f \in BAD\backslash SMALL$). Since $z_f \notin SMALL$ Lemma 23 implies that $\exists S \subseteq V_n$ such that $|S| \geq \log\log n$ and for every $z \in S$ the arc zz_f belongs to E_τ and has color c. Observe that at any iteration $\ell < f$ at most 2 vertices are removed from $V^+ \cap S$ in the case that $z_\ell \in N(N(z_f))$, and none are removed otherwise. Hence as $V^+ \cap S = \emptyset$ at the beginning of the f-th iteration we have that $2|N(N(z_f)) \cap BAD| \geq \log\log n$ which contradicts Lemmas 21 and 23.

3.7 Structure of D_c

Lemma 24. *W.h.p. $|BAD| = o(n^{1-\delta})$, for some constant $\delta > 0$.*

Proof. Recall that $p_1 = m_1/n(n-1)$. For every $v \in V_n$, (2) gives us

$$\mathbb{P}\big(v \in BAD\big) = \mathbb{P}\big(v \in B^+_1 \cup B^-_1 \cup B^+_2 \cup B^-_3\big) \leq 4\mathbb{P}(v \in B^+_1) = 4\mathbb{P}\bigg(d^+_{m_1}(v) \leq e^{-\epsilon \log n}\bigg)$$

$$\leq 12\mathbb{P}\bigg(Bin(n-1,p_1) \leq \epsilon \log n\bigg) \leq 12\exp\bigg(-0.4e^{-q\cdot 10^4}\log n\bigg) = n^{-0.4e^{-q\cdot 10^4}}.$$

At the last inequality we used (3). Hence by Markov's inequality, we have

$$\mathbb{P}\bigg(|BAD| > n^{1-0.4e^{-q\cdot 10^4}}\bigg) \leq \frac{\mathbb{E}(|BAD|)}{n^{1-0.4e^{-q\cdot 10^4}}} \leq n^{-0.09e^{-q\cdot 10^4}}. \qquad \square$$

Lemma 25. *W.h.p.* $|V_c| = n - o(n)$.

Proof. Every contraction that occurs during the execution of $HideBad$ reduces the number of vertices by one. As at most $2|BAD|$ contractions are performed, Lemma 24 gives us that w.h.p. $|V_c| \geq n - 2 \cdot n^{1-0.4e^{-q\cdot 10^4}}$. $\square$

We henceforth set $n_c := |V_c| = (1 - o(1))n$.

3.8 PHASE 1

In this section we take our first step toward proving that w.h.p. D_c has a Hamilton cycle by showing that w.h.p. there exists a matching in D_c consisting of at most $2\log n$ cycles and whose edges appear by time m_3. As usual, we proceed by implicitly conditioning on all aforementioned events proven to occur w.h.p. In the proof of Lemma 27 we are going to use the following elementary result.

Lemma 26. *W.h.p. in* D_{τ_q} *no vertex belongs to two distinct cycles of length at most 4.*

Proof. In the event that there is a vertex that belongs to two distinct cycles of length at most 4 there are $3 \leq k \leq 7$ vertices that span $k+1$ edges in D_{τ_q}. Since w.h.p. $\tau_q < 2\log n$, (2) implies that the probability of such event occurring is bounded by

$$3\sum_{k=3}^{7}\binom{n}{k}\binom{k(k-1)}{k+1}\left(\frac{2\log n}{n}\right)^{k+1} = o(1).$$

$\square$

Lemma 27. *W.h.p. every* $v \in V_c$ *has at least 6 out- and 6 in- arcs in* $E(D_c)$ *revealed during the intervals* $(m_1, m_2]$ *and* $(m_2, m_3]$ *respectively, whose other endpoint lies in* $V^* := V\backslash(BAD \cup N(BAD))$.

Here, it is imperative that we avoid $BAD \cup N(BAD)$, since those vertices have already been assigned an edge in at least one direction by the algorithm $HideBad$ from Section 3.6.

Proof. We originally defined BAD during the algorithm $DCOL$ to make sure these vertices we want to work with had many edges during the $(m_1, m_2]$ period, and the cycling between colors means a positive proportion of them obtain color c. The edges to BAD don't enjoy the cyclic colors, and the edges to $N(BAD)$ are discarded altogether even if they were in desired color c, but the estimates from Section 3.6 forbid too many of these vertices from being clustered around v.
More explicitly, let $v \in V_c$. Then by Remark (3.6.4) we have $v^+ \notin BAD$, therefore Lemma 22 gives us $v^+ \in FULL^+_{m_1}$. Now $v^+ \notin BAD \Rightarrow v^+ \notin B^+_2$, so there are at least $\epsilon \log n$ arcs v^+w, $w \in V_n$ that have been revealed after the time m_1 and before the time m_2+1. Any such arc v^+w that was *not* colored cyclically was due to $w \notin FULL^+_{m_1}$ taking priority, and hence $w \in N(v^+) \cap BAD$ by Lemma 22. So out of all the potential arcs at least $\frac{1}{q}(\epsilon \log n - |BAD \cap N(v^+)|) - 1$ have color c (see lines 6-14, 24-25 of $DCOL$), and already none of these are to BAD. Meanwhile, for $N(BAD)$, Lemma 21 immediately gives $|N(N(v^+)) \cap BAD| \leq 4e^{q \cdot 10^5}$. In addition Lemma 26 implies that $\forall w \in BAD$, $|N(v^+) \cap N(w)| \leq 2$, so any $w \in N(N(v^+))$ arose from ≤ 2 neighbours of v^+, and it follows $|N(BAD) \cap N(v^+)| \leq 2 \cdot 4e^{q \cdot 10^5}$. Hence, since $V^* := V \setminus (BAD \cup N(BAD))$, there are at least $\frac{\epsilon \log n}{q} - \frac{8}{q}e^{q \cdot 10^5} - 1 - 4e^{q \cdot 10^5} \geq 6$ arcs from v^+ to V^* in $E(D_c)$ revealed during the interval $(m_1, m_2]$.

The other part of this Lemma follows in a similar fashion (with v^-, $FULL^-_{m_1}$ and $(m_2, m_3]$ in place of v^+, $FULL^+_{m_1}$ and $(m_1, m_2]$ respectively). □

Definition 3.8.1. For $v \in V_c$ set:

$$E^+_c(v) := \{\text{the first six arcs from } v \text{ to } V^* \text{ in } E(D_c) \text{ that are revealed in } (m_1, m_2]\},$$

$$E^-_c(v) := \{\text{the first six arcs from } v \text{ to } V^* \text{ in } E(D_c) \text{ that are revealed in } (m_2, m_3]\},$$

$$E^+_c := \bigcup_{v \in V_c} E^+_c(v), \qquad E^-_c := \bigcup_{v \in V_c} E^-_c(v).$$

From Lemma 27 it follows that w.h.p. the above sets are well-defined.

Lemma 28. *W.h.p. $E^+_c \cup E^-_c$ spans a matching on V_c consisting of at most $2 \log n_c$ cycles.*

Proof. We will first show that w.h.p. $E^+_c \cup E^-_c$ spans a matching on V_c. Assume that E^+_c, E^-_c do not span a matching. Then Hall's Theorem gives us that there exists $K \subseteq V_c$ with $|K| = k \leq \frac{n_c}{2}$ that has in- or out-neighbourhood induced by E^-_c and E^+_c respectively of size $k-1$. We will examine the case of its out-neighborhood being of size $k-1$. The other case will follow in a similar fashion.

Let Y^+ be the random subgraph of D_c with edge set $E(Y^+) := E_{m_3} \setminus E^+_c$. Conditioned on $E(Y^+)$ we may assume that for every $v \in V(D_c)$, $E^+_c(v)$ has been chosen independently uniformly at random from all sets of arcs form v to $V^* \setminus \{v\}$ of size 6 that have empty intersection with $E(Y^+)$. To see this let $E(Y^+) = \{f_1, ..., f_k\}$, $h_1, ..., h_k \in [m_3]$ and for $v \in V_c$ we let $H_v \subseteq [m_3]$ such that $|H_v| = 6$. If we further conditioned on the event $\mathcal{E} = \left(\bigwedge_{i \in [k]} \{h(f_i) = h_i\} \right) \wedge \left(\bigwedge_{v \in V_c} \left\{ \{h(e) : e \in E^+_c(v)\} = H_v \right\} \right)$, in the case $\mathcal{E} \neq \emptyset$, we have

that for any $w \in V_c$ each set of arcs from w to $V^* \backslash \{w\}$ of size 6 that has empty intersection with $E(Y^+)$ has the same probability to be $E_c^+(w)$. Moreover the identity of the edges in $E_c^+(w)$ does not depend on the identity of $\{E_c^+(u) : u \in A\}$ for any $A \subseteq V_c \backslash \{w\}$.
We write $d^+_{Y^+}(v, S)$ for the number of arcs in Y^+ from v to a given $S \subseteq V_c$. Lemma 14 implies that for every $v \in V_c$, $d^+_{Y^+}(v, V_c) \leq \frac{3 \log n}{10^3 q}$. Therefore the probability of having a set $K \subseteq V_c$ that has as out neighborhood induced by E_c^+ a set $S \subseteq V^*$ with $6 \leq |S| = |K| - 1 \leq \frac{n_c}{2}$ is bounded above by

$$\sum_{k=7}^{\frac{n_c}{2}} \sum_{|K|=k} \sum_{|S|=k-1} \prod_{v \in K} \binom{k - 1 - \mathbb{I}(v \in S) - d^+_{Y^+}(v, S)}{6} \Big/ \binom{|V^*| - 1 - d^+_{Y^+}(v, V)}{6}$$

$$\leq \sum_{k=7}^{\frac{n_c}{2}} \binom{|V_c|}{k} \binom{|V^*|}{k-1} \prod_{j=1}^{k} \binom{k}{6} \Big/ \binom{|V^*| - 1 - \frac{3 \log n}{100q}}{6} \leq \sum_{k=7}^{\frac{n_c}{2}} \left(\frac{3n_c}{k}\right)^{2k} \prod_{j=1}^{k} \frac{k^6}{(1 - o(1))n_c^6}$$

$$\leq \sum_{k=7}^{\frac{n_c}{2}} \left(\frac{3^2 k^6 n_c^2}{(1 - o(1))k^2 n_c^6}\right)^k \leq \sum_{k=7}^{\frac{n_c}{2}} \left(\frac{8k^4}{(1 - o(1))n_c^4}\right)^k = o(1).$$

At the second inequality we used that Lemmas 13 & 24 imply that $n_c = |V_c| \geq |V^*| \geq |V| - |BAD| - |N(BAD)| = (1 - o(1))n = (1 - o(1))n_c$. Hence, Hall's condition fails with probability $o(1)$ and w.h.p. $E_c^+ \cup E_c^-$ spans a matching.

We proceed to prove that a random matching spanned by $E_c^+ \cup E_c^-$ consists of at most $2 \log n_c$ cycles. First let W be the number of cycles that span less than 2 vertices of V^* (i.e. 2-cycles of the form v, w with $v \in V^*$ and $w \notin V^*$). Then

$$\mathbb{P}(W \geq 1) \leq \sum_{v \in V^*} \sum_{w \notin V^*} \mathbb{P}(vw^+ \in E_c^+(w) \text{ and } w^- v \in E_c^-(w))$$

$$\leq |V^*||BAD| \left(\frac{6}{(1 + o(1))|V^*|}\right)^2 = o(1).$$

Let M be a random matching spanned by $E_c^+ \cup E_c^-$. Since w.h.p. there is no such cycle spanned by a single vertex of V^* we have that w.h.p. M induces a derangement on V^*. Finally conditioned on V^*, due to the symmetry of the edges with an endpoint in V^*, any such derangement is equally likely to occur.

Indeed let $A \subseteq V$ and consider any valid edge sequence $\mathcal{E} = e_1, ..., e_{\tau_q}$. Let ϕ_1, ϕ_2 be any two permutations on V that act as the identity on $V \backslash A$. Also let $\rho = \phi_2 \phi_1^{-1}$. Finally set $\mathcal{E}' = e'_1, ..., e'_{\tau_q}$ where for $i \in [\tau_q]$ $e_i = (u_i, w_i)$ and $e'_i = (u_i, \rho(w_i))$. Note that, provided $V \setminus A$ contains all $SMALL$ vertices, $\mathcal{E}'$ is also a valid edge sequence. Denote by $BAD_{\mathcal{E}}, V_{D,\mathcal{E}}, V^+_{\mathcal{E}}, V^-_{\mathcal{E}}, E^+_{c,\mathcal{E}}, E^-_{c,\mathcal{E}}$ and $BAD_{\mathcal{E}'}, V_{D,\mathcal{E}}, V^+_{\mathcal{E}}, V^-_{\mathcal{E}'}, E^+_{c,\mathcal{E}'}, E^-_{c,\mathcal{E}'}$ the sets BAD, V_D, V^+, V^-, E_c^+, E_c^- as defined by the sequences $\mathcal{E}$ and $\mathcal{E}'$ respectively.

First assume that $A = V^*_{\mathcal{E}}$. Then, as ρ acts on the in-vertices of arcs with in-vertex in A, we have $A = V^*_{\mathcal{E}'}$. Similarly, by considering ρ^{-1} we have $A = V^*_{\mathcal{E}}$ only if $A = V^*_{\mathcal{E}'}$. Hence $A = V^*_{\mathcal{E}}$ iff $A = V^*_{\mathcal{E}'}$. Thereafter, given that $A = V^*_{\mathcal{E}} = V^*_{\mathcal{E}'}$, we have $BAD_{\mathcal{E}} = BAD_{\mathcal{E}'}$ and

by extension, since the arcs adjacent to BAD vertices are the same and appear in the same order in both sequences, we have $V_{D,\mathcal{E}} = V_{D,\mathcal{E}'}$. Furthermore $(u,w) \in E^+_{c,\mathcal{E}}(u)$ iff $(u,\rho(w)) \in E^+_{c,\mathcal{E}'}(u)$ and $(u,w) \in E^-_{c,\mathcal{E}}(w)$ iff $(u,\rho(w)) \in E^-_{c,\mathcal{E}'}(\rho(w))$. Therefore $(u,w) \in E^+_{c,\mathcal{E}} \cup E^-_{c,\mathcal{E}}$ iff $(u,\rho(w)) \in E^+_{c,\mathcal{E}'} \cup E^-_{c,\mathcal{E}'}$. Finally, given that $A = V^*_{\mathcal{E}} = V^*_{\mathcal{E}'}$, is not hard to check that $E^+_{c,\mathcal{E}} \cup E^-_{c,\mathcal{E}}$ spans a matching on $V_{D,\mathcal{E}}$ that induces the permutation ϕ_1 on A iff $E^+_{c,\mathcal{E}'} \cup E^-_{c,\mathcal{E}'}$ spans a matching on $V_{D,\mathcal{E}'}$ that induces the permutation $\rho(\phi_1) = \phi_2$ on A. Here by induces we mean the following: if $u, u_k \in A$ and $u_1, u_2, ..., u_{k-1} \notin A$ then the matching with arcs $(u,u_1),(u_1,u_2),...,(u_{k-1},u_k)$ induces a permutation on A that sends u to u_k.

It is known (see for example [23]. [22]) that the number of cycles, in a uniform random derangement on $[|V^*|]$, consists w.h.p. of at most $2\log|V^*| \leq 2\log n_c$ cycles. Hence w.h.p. $E^+_c \cup E^-_c$ spans a matching consisting of at most $2\log n_c$ cycles. □

3.9 General Reduction

Our vertex set is V_c. Lemma 20 states that if D_c is Hamiltonian then D_{τ_q} spans a cycle of color c. Hence, in order to give a reduction of Theorem 1 to Lemma 10 we need to define digraphs $F, H, D_{n_c,p}$ on V_c such that:

i) F is a 1-factor consisting of $O(\log n_c)$ directed cycles,

ii) H has total maximum in-/out- degree $O(\log n_c)$,

iii) $D_{n_c,p}$ is a random digraph, every arc appears independently with probability $p = \Omega(\frac{\log n_c}{n_c})$

iv) w.h.p. $E(F), E(D_{n_c,p}) \subseteq D_{\tau_q}$ and all the arcs in $E(F) \cup (E(D_{n_c,p}) \setminus E(H))$ have color c.

We let F be a 1-factor spanned by $E^+_c \cup E^-_c$ consisting of at most $2\log n_c$ cycles, as provided by Lemma 28. We also let H consist of all edges that appear by time m_3. Lemma 14 implies that the maximum in/out-degree of H is $O(\log n_c)$.

For the construction of $D_{n_c,p}$ we consider the arcs appearing in $(m_3, \tau_q]$. Since

- w.h.p. $\tau_q - m_3 \geq \frac{3}{4}\log n_c$,
- w.h.p. $|BAD| = o(n_c)$,
- Every arc that appears after time m_3 and is not adjacent to BAD is colored c independently with probability $\frac{1}{q}$, and
- Every arc in D_c that has not appeared by time m_3 corresponds to exactly one arc not in D_{m_3},

we have the following (see [31]). We may couple $D_{n_c,p}$ and D_{τ_q} such that, w.h.p.:

- $E(D_{n_c,p}) \subseteq E(D_{\tau_q})$,
- Every arc spanned by V_c is present in $D_{n_c,p}$ independently with probability $p = \frac{2\log n_c}{3n_c}$, and
- If $e \in E(D_{n_c,p})$ then either e has color c or $e \in H$ (i.e. it corresponds to an arc that appears by time m_3).

By construction, $F, H, D_{n_c,p}$ satisfy the required conditions. Therefore Lemma 10 implies Theorem 1.

3.9.1 New Setup

The two next sections are given in the setup of Lemma 10 (in particular, we replace n_c by n without further comment). Thus we are given a vertex set V of size n, a 1-factor F consisting of $z = \kappa \log n$ cycles, $\kappa > 0$ and a digraph H of maximum in/out-degree $\Delta_H = O(\log n)$. Moreover we are given the random digraph $D_{n,p}$ where $p = \Omega\big(\frac{\log n}{n}\big)$.

We let ϕ be the permutation on V associated with F, i.e. $E(F) = \{(v, \phi(v)) : v \in V\}$. Furthermore we let $D^2 \sim D_{n,p'}$, $D^3 \sim D_{n,p'}$ where $p' := \frac{\xi \log n}{n} = \min\big\{\frac{p}{3}, \frac{\log n}{2n}\big\}$, for some $\xi = \xi(n) = \Omega(1)$. Since $(1-p')(1-p') \le (1-p)$, we can couple $D_{n,p}, D^2, D^3$ in such a way that $D^2 \cup D^3 \subseteq D_{n,p}$. Before proceeding we make the following observation.

Lemma 29. *W.h.p.* $\Delta(D_{n,p'}) \le 4\log n$.

Proof.

$$\mathbb{P}\big(\Delta(D_{n,p'}) \ge 4\log n\big) \le 2 \cdot n\binom{n-1}{4\log n}p'^{4\log n} \le 2n\left(\frac{en}{4\log n}\right)^{4\log n}\left(\frac{\log n}{2n}\right)^{4\log n} = o(1).\square$$

The proof of Lemma 10 is split into two parts corresponding to *Phase 2* and *Phase 3* of the algorithm in [27] that finds a Hamilton cycle in $D_{n,\frac{(1+o(1))\log n}{n}}$. Thus we refer to the first part of Lemma 10 as *Phase 2* and to the second one as *Phase 3*. As mentioned in the section "Finding a Hamilton Cycle" in *Phase 2*, we sequentially join cycles in order to create a large one consisting of $n - o(n)$ vertices. We finish the merging of all the cycles in *Phase 3*.

3.10 PHASE 2

Let $C_1, \dots, C_z$ be the cycles in F in order of decreasing size. In order to create a cycle of size at least $n - \frac{n}{\sqrt{\log n}}$ we implement the algorithm given below, denoting by (a, b) the permutation transposing a and b.

Algorithm 5 Merge Cycles

Initialize: $\phi_1 = \phi, E(\phi_1) = E(\phi), k = z$.
while *there exist* $1 \leq i < j \leq z$ *and* $a \in V(C_i), b \in V(C_j)$ *such that* $ab, \phi_1^{-1}(b)\phi_1(a) \in E(D^2)\backslash E(H)$ **do**
  $\phi_1 \leftarrow \phi_1 \circ (a, \phi_1^{-1}(b))$
  $E(\phi_1) \leftarrow \{ab, \phi_1^{-1}(b)\phi_1(a)\} \cup E(\phi_1)\backslash\{a\phi_1(a), \phi_1^{-1}(b)b\}$
  $k \leftarrow k - 1$
  Rename the cycles of ϕ_1 as $C_1, C_2, ..., C_k$ in decreasing order of size.
end
Rename the final permutation to be ϕ_2 and rename its cycles as $C'_1, C'_2, ..., C'_y$ in decreasing order of size.

Lemma 30. *W.h.p.* $|C'_1| \geq n - \frac{n}{\sqrt{\log n}}$.

Proof. Assume that after applying the algorithm above we obtain $|C'_1| < n - \frac{n}{\sqrt{\log n}}$. Set $\alpha := \max\left\{i \in [y] : \sum_{j=1}^{i} |C'_j| < n - \frac{n}{\sqrt{\log n}}\right\}$, $A := \bigcup_{i\in[\alpha]} C'_i$ (so $|A| < n - \frac{n}{\sqrt{n}}$) and $\bar{A} := V\backslash A$. As the sequence $|C'_1|, |C'_2|, ..., |C'_y|$ is decreasing, we have

$$n - \frac{n}{\sqrt{\log n}} \leq \sum_{j=1}^{\alpha+1} |C'_j| \leq 2\sum_{j=1}^{\alpha} |C'_j|.$$

Hence, $|A| = \sum_{j=1}^{i} |C'_j| \geq \frac{n}{2} - \frac{n}{2\sqrt{\log n}} \geq \frac{n}{3}$. On the other hand $|\bar{A}| = n - |A| \geq \frac{n}{\sqrt{\log n}}$. Since *Merge Cycles* ends, after performing $1 \leq k \leq z$ merges with cycles $C'_1, ...C'_y$, we have that there do not exist $1 \leq i \leq \alpha < j \leq y$ and $a \in V(C'_i)$, $b \in V(C'_j)$ such that $ab, \phi_2^{-1}(b)\phi_2(a) \in E(D^2)\backslash E(H)$. So, for every $a \in A$, $b \in \bar{A}$; either $ab \notin E(D^2)\backslash E(H)$ or $\phi_2^{-1}(b)\phi_2(a) \notin E(D^2)\backslash E(H)$. $A, \bar{A}$ define at least $n/\sqrt{\log n} \cdot n/3$ such pairs of arcs out of which at most $2|E(H)|$ have at least one edge in $E(H)$. Thus the reason that *Merge Cycles* terminates is that for each one of those, at most $\frac{n}{\sqrt{\log n}} \cdot \frac{n}{3} - 2|E(H)|$, pairs of arcs at least one does not belong to $E(D_2)$. This occurs with probability at most $(1-(p')^2)^{\frac{n}{\sqrt{\log n}}\cdot\frac{n}{3}-2|E(H)|}$ (recall $D^2 \sim D_{n,p'}$).

Merge Cycles performs some number $k \leq z := \kappa \log n$ merges. Each such merge is uniquely determined by one of its arcs (i.e. either ab or $\phi_1^{-1}(b)\phi_1(a)$). Hence at every execution of the while loop of *Merge Cycles* there are at most $n(n-1)$ possible merges available. Therefore for $0 \leq k \leq z$ there are most $[n(n-1)]^k$ sequences of k merges that *Merge Cycles* may perform. Any of those sequences may take place only if the corresponding $2k$ arcs lie in $E(D^2) \setminus E(H)$, so any sequence occurs with probability at most $(p')^{2k}$. Thus, by considering the number of merges k, all the possible sequences of k merges that *Merge Cycles* may perform, the probability that a given sequence the related arcs lie in $E(D^2)$ and the probability of *Merge Cycles* terminating due to lack of additional edges after performing this exact sequence of k

merges, we have

$$\begin{aligned}\mathbb{P}\left(|C_1'| < n - \frac{n}{\sqrt{\log n}}\right) &= \sum_{k=0}^{z} \left[n(n-1)\right]^k (p')^{2k} (1-(p')^2)^{\frac{n}{\sqrt{\log n}}\cdot\frac{n}{3} - 2|E(H)|} \\ &\leq \sum_{k=0}^{\kappa \log n} (\xi \log n)^{2k} \cdot \exp\left\{-\frac{\xi^2 \log^2 n}{n^2}\left[\frac{n}{\sqrt{\log n}}\cdot\frac{n}{3} - 2n\Delta_H\right]\right\} \\ &\leq (\kappa \log n + 1)\cdot(\xi \log n)^{2\kappa \log n} \cdot \exp\left(-(1+o(1))\xi^2 \log^{1.5} n\right) \\ &= o(1). \end{aligned}$$

□

3.11 PHASE 3

With high probability we inherit from *Phase 2* a permutation ϕ_2 consisting of y cycles, $C_1', ..., , C_y'$ such that $|C_1'| \geq |C_2'| \geq ... \geq |C_y'|$, $|C_1'| \geq n - \frac{n}{\sqrt{\log n}}$ and $y \leq \kappa \log n$. We also inherit the edges $E(\phi_2)$ associated with the permutation ϕ_2. We will use the edges in $E(D^3)$, recalling $D^3 \sim D_{n,p'}$, in order to merge one by one all the cycles with C_1'. At iteration i of *Phase 3* we merge C_i' with the cycle $C(i-1)$. $C(i-1)$ is the output of iteration $i-1$ of *Phase 3* and it spans $C_1', ..., C_{i-1}'$. The merging of C_i' with $C(i-1)$ is performed by *FindCycle*$(C(i-1), C_i',$ *outcome*$)$.

To merge the two cycles we start by finding arcs in $E(D^3) \setminus E(H)$ from C_i' to $C(i-1)$. For every such arc we create a di-path that spans $V(C_i') \cup V(C(i-1))$ and uses the edges of the two cycles in addition to the selected arc. We let the set of those di-paths be $\mathcal{P}_0^i$- we will now use the Pósa rotations to grow $\mathcal{P}_0^i$ exponentially. Precisely, at iteration t of *FindCycle*$(C(i-1), C_i',$ *outcome*$)$ we are given a set of di-paths that spans $V(C_i') \cup V(C(i-1))$ which we denote by $\mathcal{P}_{t-1}^i$. For every di-path $p_r \in \mathcal{P}_{t-1}^i$ we generate every possible di-path that can be obtained from p_r by a single double rotation (i.e. a two arc exchange; see Figure 2/ Section 5) with the sole condition being that the two new arcs should belong to $E(D^3) \setminus E(H)$. The new di-paths generated at iteration t are added to $\mathcal{P}_{t-1}^i$ to create $\mathcal{P}_t^i$. We grow this collection of paths $T = \frac{\log n}{\log \log n}$ times. By this point, there are so many di-paths in $\mathcal{P}_T^i$ that a *constant proportion* of all vertices have become an endpoint, and so we have a good chance to close at least one into a cycle using another arc in $E(D^3) \setminus E(H)$.

Once more, we proceed by implicitly conditioning on all aforementioned events that are proven to occur w.h.p.

Algorithm 6 Phase 3

$C(1) = C'_1$
for $i = 2, 3, ..., y$ **do**
  outcome $\leftarrow$ failure
  suppose $C'_i = (x_{i,1}, x_{i,2}, ..., x_{i,n_i})$
  Execute FindCycle$(C(i-1), C'_i,$ outcome)
  if *outcome = failure* **then**
    Terminate *Phase 3*
  end
end

Algorithm 7 FindCycle$(C(i-1), C'_i,$ outcome)

Suppose $C(i-1) = (y_1, y_2, ..., y_\gamma)$.
Set
$\mathcal{P}^i_0 := \{(x_{i,1}, x_{i,2}, ..., x_{i,n_i}, y_j, y_{j+1}, ..., y_\gamma, y_1, ..., y_{j-1}) : j \in [\gamma]$ and $x_{i,n_i}y_j \in E(D^3) \setminus E(H)\}$.
for $t = 1, ..., \lfloor \frac{\log n}{\log\log n} \rfloor$ **do**
  Suppose $\mathcal{P}^i_{t-1} = \{p_1, p_2, ..., p_s\}$;
  $\mathcal{P}^i_t := \mathcal{P}^i_{t-1}$
  for $r = 1, ..., s$ **do**
    Suppose $p_r = (u_1, u_2, ..., u_\ell)$
    For all (a, b) such that $a < b$ and $(u_\ell, u_a), (u_{a-1}, u_b) \in E(D^3) \setminus E(H)$ set:
    $\mathcal{P}^i_t \leftarrow \mathcal{P}^i_t \cup \{(u_1, u_2, ..., u_{a-1}, u_b, u_{b+1}, ..., u_\ell, u_a, u_{a+1}, ..., u_{b-1})\}$
  end
end
Suppose $\mathcal{P}^i_{\lfloor \frac{\log n}{\log\log n} \rfloor} = \{p_1, p_2, ..., p_d\}$.
for $k = 1, ..., d$ **do**
  Suppose $p_k = (w_1, w_2, ..., w_\zeta)$
  if $(w_\zeta, w_1) \in E(D^3) \setminus E(H)$ **then**
    $C(i) = (w_1, w_2, ..., w_\zeta, w_1)$
    outcome $\leftarrow$ success
    Terminate FindCycle$(C(i-1), C'_i,$ outcome)
  end
end

With $n_1 = |C'_1|$ let $C'_1 = (v_1, v_2, ..., v_{n_1}, v_1)$. Partition C'_1 into $\mu_1 := \lceil \log^2 n / \log\log\log n \rceil$ intervals $A_1, A_2, ...$ of size $\lceil |C'_1|/\mu_1 \rceil$ or $\lfloor |C'_1|/\mu_1 \rfloor$, namely $A_i = \{v_{r_{i-1}+1}, v_{r_{i-1}+2}, ..., v_{r_i}\}$ for some $0 = r_0 < r_1 < r_2 < ... < r_{\mu_1} = n_1$. For $I \subseteq [\mu_1]$ let $A_I := \bigcup_{i \in I} A_i$, $n_I := |A_I|$ and

$B_I := \{v \in V(C_1') : |\{u \in A_I : (v,u) \in E(D^3) \setminus E(H)\}| \le \xi \log n/20\}$ be the set of all vertices with much fewer than the expected number of out-neighbours to the I-intervals in $D^{(3)} \setminus H$.

Lemma 31. *W.h.p for all $I \subseteq [\mu_1]$ with $|I| = \lfloor \mu_1/10 \rfloor$ we have that $|B_I| \le n^{1-\frac{\xi}{100}}$.*

Proof. For a fixed such I we have $n_I = \sum_{l \in I} |A_l| \ge |I| \lfloor |C_1'|/\mu_1 \rfloor \ge \big(\frac{\mu_1}{10} - 1\big)\big(\frac{|C_1|}{\mu_1} - 1\big)$. Therefore as $n_1 = |C_1'| = \left(1 - \frac{1}{\sqrt{\log n}}\right) n$ we get that $n_I = (1 + o(1))0.1n$. Moreover, for any vertex $v \in V$ there are at most $\Delta_H = O(\log n)$ arcs in $E(H)$ from v to A_I. Hence, for fixed k:

$$\begin{aligned}
\mathbb{P}\big(|B_I| \ge k\big) &\le \binom{n_1}{k} \mathbb{P}\left[Bin\left(n_I - \Delta_H, \frac{\xi \log n}{n}\right) \le \frac{\xi \log n}{20} \right]^k \\
&\le \left(\frac{en}{k}\right)^k \left[\exp\left(-(1+o(1)) \frac{0.5^2}{2} \frac{\xi \log n}{10} \right) \right]^k \\
&= \left(\frac{e}{k} n^{1 - \frac{(1+o(1))\xi}{80}} \right)^k \le \left(\frac{e}{k} n^{1 - \frac{\xi}{90}} \right)^k.
\end{aligned}$$

At the 2nd inequality we used the Chernoff bounds (3). Thus, with $k = n^{1-\frac{\xi}{100}}$ we have

$$\begin{aligned}
\mathbb{P}\big(\exists I \subseteq [\mu_1] : |I| = \lfloor \mu_1/10 \rfloor; |B_I| \ge n^{1-\frac{\xi}{100}}\big) &\le \binom{\mu_1}{\lfloor \mu_1/10 \rfloor} \left(\frac{e}{n^{1-\frac{\xi}{100}}} n^{1-\frac{\xi}{90}} \right)^{n^{1-\frac{\xi}{100}}} \\
&\le 2^{\mu_1} \left(e n^{-\frac{\xi}{1000}} \right)^{n^{1-\frac{\xi}{100}}} = o(n^{-1}). \qquad \square
\end{aligned}$$

Next, let $\mu_2 := \lceil \frac{\log n}{\log\log\log\log n} \rceil$.

Lemma 32. *W.h.p. for every $v \in V$ and every $I \subseteq [\mu_1]$ with $|I| = \lfloor \mu_1/10 \rfloor$, we have $|\{b \in B_I : v\phi_2(b) \in E(D^3)\}| < \mu_2$.*

Proof. For fixed v, I, and $B = \{b_1, b_2, ..., b_{\mu_2}\}$, the probability that every $v\phi_2(b_i) \in E(D^3)$ and $B \subseteq B_I$ is bounded by

$$\left(\frac{\xi \log n}{n}\right)^{\mu_2} \cdot \mathbb{P}\left[Bin\left(n_I - \Delta_H - \mathbb{I}(v \in A_I), \frac{\xi}{\log n}\right) \le \frac{\xi \log n}{20} \right]^{\mu_2} \le \left(\frac{\xi \log n}{n}\right)^{\mu_2} \cdot n^{-\frac{\xi \mu_2}{90}}.$$

Therefore,

$$\mathbb{P}(\exists v, I, B \text{ as above}) \le n 2^{\mu_1} \binom{n}{\mu_2} \left(\frac{\xi \log n}{n}\right)^{\mu_2} \cdot n^{-\frac{\xi\mu_2}{90}} \le n 2^{\mu_1} \left(\frac{en}{\mu_2}\right)^{\mu_2} \left(\frac{\xi \log n}{n}\right)^{\mu_2} \cdot n^{-\frac{\xi\mu_2}{90}}$$

$$\leq \exp\left\{ \log n + \mu_1 \log 2 + \mu_2 \log\left(\frac{e\xi \log n}{\mu_2}\right) - \frac{\xi\mu_2}{90}\log n \right\}$$
$$\leq \exp\left\{\Theta(\mu_1 - \mu_2 \log n)\right\} = o(1).$$

□

Lemma 33. *Let $0 < \alpha < 1$ be fixed. Then w.h.p. there do not exist $A, B \subseteq V(C_1')$ satisfying all 3 of the following:*

i) $|A| \leq \alpha_0 = \alpha e^{-3} n/\log n$,

ii) $|B| \leq \alpha |A| \log n/2$

iii) $|\{(u,v) \in E(D^3) : u \in A, v \in B\}| \geq \alpha |A| \log n$.

Proof. Observe that if there exist sets A, B satisfying conditions i-iii we may extend B, by adding to it any vertices of $V(C_1')$, to a set B' of size $\alpha|A|\log n/2$ such that the sets A, B' also satisfy conditions i-iii. Hence, if we let $\mathcal{F}$ be the event that there exist sets A, B satisfying conditions i-iii, then as $|V(C_1')| \leq n$,

$$\mathbb{P}(\mathcal{F}) \leq \sum_{k=1}^{\alpha_0} \sum_{\substack{A,B\subseteq V(C_1'):\\ |A|=k, |B|=\alpha k \log n/2}} \sum_{\substack{E\subseteq A\times B:\\ |E|=\alpha k \log n}} \left(\frac{\xi \log n}{n}\right)^{\alpha k \log n}$$
$$\leq \sum_{k=1}^{\alpha_0} \binom{n}{k}\binom{n}{\alpha k \log n/2}\binom{k\cdot \alpha k \log n/2}{\alpha k \log n} \cdot \left(\frac{\xi \log n}{n}\right)^{\alpha k \log n}$$
$$\leq \sum_{k=1}^{\alpha_0} \left\{ \frac{en}{k}\left[\frac{2en}{\alpha k \log n}\left(\frac{ek}{2}\right)^2 \left(\frac{\xi \log n}{n}\right)^2\right]^{\alpha \log n/2}\right\}^k$$
$$\leq \sum_{k=1}^{\alpha_0} \left[\frac{en}{k}\left(\frac{ke^3\xi \log n}{2\alpha n}\right)^{\alpha \log n/2}\right]^k = o(1).$$

At the last line we used that $\xi \leq \frac{1}{2}$ and that $k \leq \alpha e^{-3} n/\log n$. □

We say that iteration i of *Phase 3* is a success if *FindCycle*$(C(i-1), C_i', outcome)$ merges $C(i-1)$ with C_i'. To show that *Phase 3* is successful it is enough to show that for $i \in [y]$, conditioned on iteration $i-1$ of the algorithm being a success (i.e. *Findcycle* defines $C(i-1)$), iteration i is not a success with probability $o(\frac{1}{\log n})$ (there are $O(\log n)$ cycles to be merged). Henceforth we implicitly condition on the statements of the previous three Lemmas.

The following three definitions will be of high significance for the rest of this section.

Definition 3.11.1. For $I \subseteq [\mu_1]$ set $cl(A_I) := \{e \in E(C_1') : |e \cap V(A_I)| \geq 1\}$, the edges of the large cycle corresponding to the collection of intervals I (together with their boundaries).

Definition 3.11.2. We say that a path $P = (v_1, v_2, ..., v_p)$ is *good* if $\exists I \subseteq [\mu_1]$ with $|I| = \lfloor \mu_1/10 \rfloor$ and $r < s \leq \frac{p}{2}$ such that $s - r \leq \frac{p}{9}$, $cl(A_I) \subseteq \{v_j v_{j+1} : r \leq j < s\}$ and $v_p \notin B_I$ (recall $v_p \notin B_I$ if there are more than $\frac{\xi \log n}{20}$ arcs in $E(D^3) \setminus E(H)$ from v_p to A_I).

Definition 3.11.3. For a subgraph $S \subseteq C(i-1)$, set $J_S := \Big(\bigcup_{k=2}^{i} V(C'_k) \Big) \cup \Big(\bigcup_{\ell \in F_S} A_\ell \Big)$ for $F_S := \{\ell \in [\mu_1] : cl(A_\ell) \not\subseteq E(S)\}$. This J_S should be considered as a set of junk: we want to restrict ourselves to only trying more rotations using the intervals $\ell \in [\mu_1]$ preserved from the original large cycle C'_1 which are still wholly contained in S (i.e. were not broken by a previous rotation). Certainly therefore we want to avoid any vertices leftover from the smaller cycles C'_k that have previously been merged.

Lemma 34. *Suppose S is a good path that satisfies $S \in \mathcal{P}_t^i$ for some $0 \leq t \leq \frac{\log n}{\log \log n}$. Then $|J_S| = o(n)$.*

Proof. To merge $C(i-1)$ with C'_i, we start by joining the two cycles using an edge in $E(D^3) \setminus E(H)$, then delete an edge from each cycle to create a path. Thereafter, in order to create a new path from a given one, we perform double rotations (defined in section Finding Hamilton cycles - Overview). Every double rotation involves removing two edges from the current path and adding two edges from $E(D^3) \setminus E(H)$. As *FindCycle*$(\cdot)$ performs $\leq \frac{\log n}{\log \log n}$ rounds of double rotations, $|E(C(i-1)) \setminus E(S)| \leq 1 + 2 \cdot \frac{\log n}{\log \log n}$. Similarly, $|E(C(k-1)) \setminus E(C(k))| \leq 1 + 2 \cdot \frac{\log n}{\log \log n}$ for every $2 \leq k < i$. Thus, as $i \leq \log n$, we have

$$|F_S| \leq 2|E(C'_1) \setminus E(S)| = 2|E(C(1)) \setminus E(S)| \leq 4 \log n \cdot \left(1 + 2 \cdot \frac{\log n}{\log \log n}\right) = o(\mu_1).$$

(At the first inequality, we used that each removed $e \in E(C'_1)$ was in ≤ 2 of the $cl(A_\ell)$'s). Therefore,

$$|J_S| \leq \sum_{k=2}^{i} |V(C'_k)| + \sum_{\ell \in F_S} |A_\ell| \leq o(n) + o(\mu_1) \cdot (n/\mu_1 + 1) = o(n). \qquad \square$$

Definition 3.11.4. Let $i \in [y]$ and $x \in V(C'_i)$. For $t \leq \frac{\log n}{\log \log n}$ we define $\mathcal{GP}_t^i$ to be the set of all good paths that are contained in $\mathcal{P}_t^i$. Furthermore let $ENDG_t^i$ be the set of endpoints of paths in $\mathcal{GP}_t^i$.

Lemma 35. *For $i \in [y]$, conditioned on iteration $i-1$ being a success, $\mathbb{P}(\mathcal{GP}_t^i \neq \emptyset) \geq 1 - o(n^{-\frac{\xi}{2}})$.*

Proof. Let $C(i-1) = \{u_1, u_2, ..., u_\gamma, u_1\}$. Partition $C(i-1)$ into 9 blocks $S_1, S_2, ..., S_9$ which are subpaths of near-equal length by setting, for each $\ell \in [9]$, $S_\ell := \{u_{\lfloor \frac{\ell-1}{9} \cdot \gamma \rfloor + 1}, ..., u_{\lfloor \frac{\ell}{9} \cdot \gamma \rfloor}\}$. Note every $|J_{S_\ell} \cap S_\ell| \leq |J_{C(i-1)}| + 2 = o(n)$, so

$$\sum_{\substack{i \in [\mu_1] \\ cl(A_i) \subseteq E(S_\ell)}} |A_i| = |S_\ell \setminus J_{S_\ell}| = |S_\ell| - o(n) \geq \left| \frac{C'_1}{9} \right| - 1 - o(n) = \big(1 - o(1)\big) \frac{n}{9}.$$

For every $\ell \in [9]$, let $I_\ell{}' = \{i \in [\mu_1] : cl(A_i) \subseteq E(S_\ell)\}$. (3.11) implies that $|I'_\ell| \geq \mu_1/10$. Thus we may let $I_\ell \subseteq I'_\ell$ be the set of the $\lfloor \mu_1/10 \rfloor$ smallest elements of I'_ℓ.

Recall the notation $C'_i = \{x_{i,1}x_{i,2}, ..., x_{i,n_i}, x_{i,1}\}$. $\mathcal{GP}^i_0$ in non-empty if there exists an arc $(x_{i,n_i}, u_a) \in E(D^3) \setminus E(H)$ for some $a \in [\gamma]$ such that

(i) $u_a \in A_{I_\ell}$ for some $\ell \in [9]$, and

(ii) $\phi_2^{-1}(u_a) \notin B_{I_1} \cup B_{I_2} \cup ... \cup B_{I_9}$.

Indeed let $P = \{x_{i,1}, ..., x_{i,n_i}, u_a, u_{a+1}, ..., u_\gamma, u_1, ..., u_{a-1}\}$ be such a path. Observe that $\exists j \in [9]$ such that S_j defined above is found in the interior of the first half of P (here we only needed that $C(i-1)$ was split into at least 5 blocks). In addition S_j consists of $\frac{n}{9} - o(n)$ consecutive vertices in $C(i-1)$ hence in P. Thus since $I_j \subseteq I'_j \subsetneq S_j$, $I := I_j$ is a witness to the goodness of path P. Furthermore $u_a \in A_{I_\ell}$ implies that $(\phi_2^{-1}(u_a), u_a) \in E(C(i-1))$ and therefore $\phi_2^{-1}(u_a) = u_{a-1}$. Finally since the endpoint of P, $u_{a-1} = \phi_2^{-1}(u_a) \notin B_{I_1} \cup B_{I_2} \cup ... \cup B_{I_9}$ we have that all the conditions for P to be good are met.

Lemma 31 implies that the number of vertices u_a satisfying both conditions (i) and (ii) is $(1+o(1))0.9n$. Since we do not examine the arcs in $\{x_{i,n_1}\} \times V(C'_1)$ that are found in $E(D^3)$ until we execute the i-th iteration of *Phase 3*, we have that any arc in $\{x_{i,n_1}\} \times V\big(\bigcup_{i\in[\ell]} A_{I_\ell}\big)$ not found in $E(H)$ belongs to $E(D^3)$ with probability $p' = \frac{\xi \log n}{n}$. Pause for a moment to recall that every vertex has at most $\Delta_H = O(\log n)$ out-arcs in $E(H)$ that we cannot use. Thus, given that iteration $i-1$ is a success, the probability of the event $\{\mathcal{GP}^i_0 = \emptyset\}$ is bounded above by

$$\mathbb{P}\Big\{ Bin\big[((1+o(1))0.9n - \Delta_H, p'\big] = 0 \Big\} \leq (1-p')^{(1+o(1))0.9n} \leq e^{-(1+o(1))0.9p'n} = o(n^{-\frac{\xi}{2}}).\square$$

We will use the endpoints of good paths in order to lower bound the number of distinct endpoints of paths created at some iteration of *Phase 3*. The advantage of good paths is that their endpoints have many arcs towards earlier vertices of the path, whose predecessors in turn have many arcs to vertices nearer the end of the path. Hence, we expect the number of paths originating from a specific good path after an iteration of *Phase 3* to be large. Note that for any $i \in [y]$ all the paths that are constructed during $FindCycle(C(i-1), C'_i, outcome)$ have the same starting point, namely $x_{i,1}$.

Lemma 36. *Let $i \in [y]$ be such that $\mathcal{GP}^i_t \neq \emptyset$. Then, w.h.p. for $t \leq \frac{\log n}{\log\log n} - 1$,*

$$|ENDG^i_t| \leq \frac{\xi n}{84e^3 \log^2 n} \quad \textit{implies} \quad \left(\frac{\xi \log n}{42}\right)^2 |ENDG^i_t| \leq |ENDG^i_{t+1}|.$$

Proof. For $t \leq \frac{\log n}{\log\log n} - 1$ let $P = (u_1, u_2, ..., u_p) \in \mathcal{GP}^i_t$ and r_P, s_P, I_P be as in the definition of a good path. Partition P into 9 sub-paths $S_{1,P}, S_{2,P}, ..., S_{9,P}$ containing $A_{I_{1,P}}, A_{I_{2,P}}, ...A_{I_{9,P}}$

as is done earlier in Lemma 35. Set

$$H_1(P) = \{u_j \in P : u_p u_j \in E(D^3) \setminus E(H), u_j \in A_{I_P} \text{ and } u_{j-1} \notin B_{I_{9,P}}\}$$

and

$$H_2(P) = \{u_{j-1} : u_j \in H_1(P)\}.$$

Since P is a good path we have that $u_p \notin B_{I_P}$. Therefore u_p has at least $\frac{\xi \log n}{20}$ neighbours in A_{I_P} out of which at most μ_2 have their predecessor in $B_{I_{9,P}}$ (see Lemma 32). Hence we have that

$$|H_2(P)| = |H_1(P)| \geq \frac{\xi \log n}{20} - \mu_2 \geq \frac{\xi \log n}{21}.$$

Furthermore, if $r_P < \frac{p}{9} + 1$ for each $u \in H_2(P)$ set,

$$H_3(P,u) = \{u_\ell \in P : uu_\ell \in E(D^3) \setminus E(H), u_\ell \in A_{I_{9,P}} \text{ and } u_{\ell-1} \notin B_{I_{3,P}}\}.$$

Otherwise, set

$$H_3(P,u) = \{u_\ell \in P : uu_\ell \in E(D^3) \setminus E(H), u_\ell \in A_{I_{9,P}} \text{ and } u_{\ell-1} \notin B_{I_{1,P}}\}.$$

Finally in both of the above cases set

$$H_4(P,u) = \{u_{\ell-1} : u_\ell \in H_3(P,u)\}.$$

As before, from $H_2(P) \cap B_{I_{9,P}} = \emptyset$ together with Lemma 32 we have that, for all $u \in H_2(P)$,

$$|H_4(P,u)| = |H_3(P,u)| \geq \frac{\xi \log n}{20} - \mu_2 \geq \frac{\xi \log n}{21}.$$

Finally for $k \in \{1,2\}$ and $m \in \{3,4\}$ set,

$$\mathcal{H}_k := \bigcup_{P \in \mathcal{GP}_t^i} H_k(P) \qquad \mathcal{H}_m := \bigcup_{P \in \mathcal{GP}_t^i} \left\{ \bigcup_{v \in H_2(P)} H_m(P,v) \right\}.$$

Claim: $\mathcal{H}_4 \subseteq ENDG_{t+1}^i$.

Proof of the claim: Indeed, suppose that $r_P < \frac{p}{9} + 1$ and $u_{k-1} \in \mathcal{H}_4$, i.e. there are j and k such that

$$u_p u_j, u_{j-1}u_k \in F_c^3, \quad u_j \in A_{I_P}, \quad u_k \in A_{I_{9,P}}, \quad u_{j-1} \notin B_{I_{9,P}} \text{ and } u_{k-1} \notin B_{I_{3,P}}.$$

Then, $r_P \leq j \leq s_P \leq \frac{p}{2} \leq k$ and hence a double rotation on P using the edges $u_p u_j, u_{j-1}u_k$ will result in the path $P' = (u_1, u_2, ..u_{j-1}u_k, u_{k+1}, ..., u_p, u_j, u_{j+1}, ..., u_{k-1})$. So in showing that $u_{k-1} \in ENDG_{t+1}^i$ it suffices to show that P' is a good path with $I_{P'} = I_{3,P}$. To see this first note $u_{k-1} \notin B_{I_{3,P}}$. Secondly $cl(A_{I_{3,P}}) \subseteq P'$ as $cl(A_{I_{3,P}}) \subseteq P$ and no edge of $cl(A_{I_{3,P}})$

was deleted in a double rotation. Thirdly if we let r', s' to be respectively the smallest and largest indices of vertices in $A_{I_{3,P}}(= A_{I_{P'}})$ in the path P then $(s'+1)-(r'-1) \leq \frac{|P'|}{9}(= \frac{p}{9})$ as $cl(A_{I_{3,P}}) \subseteq E(S_{3,P})$. This implies that $cl(A_{I_{P'}}) \subseteq \{u_j u_{j+1} : (r'-1)+(p-k+1) \leq j < (s'+1)+(p-k+1)\}$ and that $[(s'+1)-(p-k+1)]-[(r'-1)-(p-k+1)] \leq \frac{p}{9}$. Finally as $u_k \in A_{I_{9,P}}$ and $u_{s'} \in A_{I_{3,P}}$, we get that $p-k \leq \frac{p}{9}$ and $(s'+1) \leq \frac{p}{3}$. Hence $(s'+1)+(p-k+1) < \frac{p}{2}$.

In the case that $r_P > \frac{p}{9}$ and $u_{k-1} \in \mathcal{H}_4$, the goodness of p' (now with $I_{P'} = I_{1,P}$) follows from the same reasoning with the only difference that the vertices in $A_{I_{P'}}$ hold the same positions in both paths. Thus in both cases P' is good, proving the claim.

Suppose that $|ENDG_t^i| \leq \frac{\xi n}{84e^3 \log^2 n}$. To make sure that the endpoints of good paths in $\mathcal{GP}_{t+1}^i$ do not coincide too often we apply Lemma 33 with $\alpha = \frac{\xi}{21}, A = ENDG_t^i, B = \mathcal{H}_1$. Recall for every good path there are at least $\frac{\xi \log n}{21}$ edges in $E(D^3)\backslash E(H)$ from its endpoint that lie in A to vertices in $B = \mathcal{H}_1$. So by summing over a maximal set of paths with distinct endpoints we get that there are at least $\frac{\xi}{21}|A|\log n$ arcs from A to B. Hence as $|A| \leq \frac{\xi n}{84e^3 \log^2 n} \leq \alpha e^{-3} n/\log n$ in the Lemma 33 condition ii) must not be satisfied. Moreover Lemma 29 implies that w.h.p. there are at most $\Delta(D^3)|A| \leq 4\log n|A|$ arcs from A to B. Therefore,

$$\frac{\xi \log n}{42}|ENDG_t^i| \leq |\mathcal{H}_1| = |\mathcal{H}_2| \leq 4\log n|ENDG_t^i| \leq \frac{\xi n}{21e^3 \log n}.$$

Similarly by reapplying Lemma 33 with $\alpha = \frac{\xi}{21}, A = \mathcal{H}_2, B = \mathcal{H}_3$ we have that,

$$\left(\frac{\xi \log n}{42}\right)^2 |ENDG_t^i| \leq \frac{\xi \log n}{42}|\mathcal{H}_2| \leq |\mathcal{H}_3| = |\mathcal{H}_4| \leq |ENDG_{t+1}^i|. \qquad \square$$

Summarising, the two last lemmas give us that conditioned on phase $i-1$ being a success, $1 \leq |ENDG_0^i|$ with probability at least $1 - o(n^{-\frac{\xi}{2}})$. Furthermore since $n \leq \left(\frac{\xi \log n}{42}\right)^{\frac{1.8\log n}{\log\log n}}$ the integer $t_f := \min\left\{j : \left(\frac{\xi \log n}{42}\right)^{2j} \geq \frac{\xi n}{84e^3 \log^2 n}\right\}$ is less than $\frac{0.9\log n}{\log\log n}$ and satisfies, due to Lemma 36, $|ENDG_{t_f}^i| \geq \frac{\xi n}{84e^3 \log^2 n}$. Thus by applying the same argument as in the previous lemma to a subset F of $ENDG_{t_f}^i$ of size $\frac{\xi n}{84e^3 \log^2 n}$ and to the set of paths in $\mathcal{GP}_{t_f}^i$ with endpoints in F we have that

$$\beta n = \left(\frac{\xi \log n}{42}\right)^2 \cdot \frac{\xi n}{84e^3 \log^2 n} \leq |ENDG_{t_f+1}^i(v)|$$

for some constant $\beta > 0$. Recall that all the paths in $\mathcal{GP}_{t_f+1}^i$ start from the same vertex $x_{1,i} \in V(C_i')$ and that $\mathcal{GP}_{t_f+1}^i \subseteq \mathcal{P}_{\lfloor \frac{\log n}{\log\log n} \rfloor}^i$. Since we do not examine the arcs going into $x_{i,1}$ until the very end of the i-th iteration of *Phase 3*, *after* conditioning on iteration $i-1$ of *Phase 3* being a success every arc in $V(C_1') \times \{x_{i,1}\}\backslash E(H)$ still belongs to $E(D^3)$ with probability p' . Hence, the probability of iteration i of *Phase 3* not being a success conditioned on iteration $i-1$ is bounded by

$$o\big(n^{-\frac{\xi}{2}}\big) + \mathbb{P}\big[Bin(\beta n - \Delta_H, p') = 0\big] \leq o\big(n^{-\frac{\xi}{2}}\big) + (1-p')^{\beta n - O(\log n)} = o(n^{-\epsilon}),$$

for some $\epsilon > 0$. As we merge cycles at most $y \leq \kappa \log n$ times, *Phase 3* succeeds in merging all the cycles into one with probability $1 - o(n^{-\epsilon} \cdot \kappa \log n) = 1 - o(1)$. Finally observe that during phases 2 and 3 we use edges only in $(E(D^2) \cup E(D^3))\backslash E(H)$ which completes the proof of Lemma 10.

Chapter 4

An Inverted Turán Problem

4.1 Introduction and Motivation

For a graph G and a family of graphs $\mathcal{H}$, the extremal number of $\mathcal{H}$ in G is defined to be

$$\mathrm{ex}(G, \mathcal{H}) = \max\{|E(F)| : F \subseteq G \text{ and } H \not\subseteq F \text{ for any } H \in \mathcal{H}\}.$$

When the family consists only of a single graph, $\mathrm{ex}(G, H)$ is used in place of $\mathrm{ex}(G, \{H\})$.

A typical example of this is when $\mathcal{H} = \{C_3, C_4, C_5, \dots\}$ is the collection of all cycles, in which case the extremal number is simply the *graphic matroid rank* of G, an important graph parameter in its own right.

The Turán problem, one of the cornerstones of extremal graph theory concerns the behavior of $\mathrm{ex}(K_n, H)$ for a fixed H when n is large. The first result along these lines is a theorem of Mantel (see, for instance [11]) which states that $\mathrm{ex}(K_n, K_3) = \lfloor n^2/4 \rfloor$. Turán [43] obtained a version for K_t in place of K_3, in particular obtaining $\mathrm{ex}(K_n, K_t) = \left(1 - \frac{1}{t-1} + o(1)\right)\frac{n^2}{2}$ where $o(1) \to 0$ as $n \to \infty$. In a similar spirit, the Erdős-Stone Theorem [20] states that if $\chi = \chi(H)$ is the chromatic number of H, then $\mathrm{ex}(K_n, H) = \left(1 - \frac{1}{\chi-1} + o(1)\right)\frac{n^2}{2}$. The Erdős-Stone Theorem asymptotically answers the Turán problem, except when H is bipartite, in which case the bound becomes $o(n^2)$. In this situation, known as the degenerate case, the asymptotic behavior of very few graphs is known and is an active area of research (c.f. [28]).

Most approaches in the case of a bipartite graph instead ask about $\mathrm{ex}(K_{n,n}, H)$, which is known as the Zarankiewicz problem [46]. This is often seen as a more natural question and provides bounds on the Turán problem as $\frac{1}{2}\mathrm{ex}(K_n, H) \leq \mathrm{ex}(K_{n/2,n/2}, H) \leq \mathrm{ex}(K_n, H)$ for bipartite H. In the special case of $H = C_4$, the incidence graphs showing tightness for the Zarankiewicz problem were spotted a few years before polarity graphs showing tightness for the Turán problem (see [28, Section 3]).

With this in mind, we set out to explore a framework in which to ask: what is the most "natural" or "best" host graph for a fixed family of graphs? This suggests optimizing a

particular monotone graph parameter over all host graphs G where $\mathrm{ex}(G, \mathcal{H})$ is bounded, the simplest of which is just the edge count. Thus we define the following extremal function for $\mathcal{H}$:

$$\mathcal{E}_k(\mathcal{H}) := \sup\{|E(G)| : \mathrm{ex}(G, \mathcal{H}) < k\}.$$

In other words, for a family $\mathcal{H}$, we would like to determine the host graph G with the most edges such that any k edges from G contain some copy of $H \in \mathcal{H}$. In other words, G is best at "forcing" a copy of some $H \in \mathcal{H}$. When the family consists only of a single graph, we write $\mathcal{E}_k(H)$ in place of $\mathcal{E}_k(\{H\})$. Note that it is necessary to consider the supremum here as $\mathcal{E}_k(\mathcal{H})$ may be infinite. In particular, $\mathcal{E}_k(K_{1,t}) = \mathcal{E}_k(tK_2) = \infty$ for $k \geq t$ as for any $s \geq t$, $\mathrm{ex}(K_{1,s}, K_{1,t}) = t - 1 = \mathrm{ex}(sK_2, tK_2)$, despite both host graphs having s edges. However, we will later show that stars and matchings classify all families having $\mathcal{E}_k(\mathcal{H}) = \infty$.

In a similar fashion to the original Turán problem, this chapter considers two questions:

- What are the asymptotics of $\mathcal{E}_k(\mathcal{H})$?
- When $\mathcal{E}_k(\mathcal{H})$ can be determined precisely, which host graphs G attain $|E(G)| = \mathcal{E}_k(\mathcal{H})$?

On the one hand, we will show that for nonbipartite H, this question behaves more or less as one might expect. For example, the following theorem is close in spirit to the Erdős-Stone Theorem:

Theorem 4.1.1. *If $\mathcal{H}$ is a family of graphs with $\rho = \min\{\chi(H) : H \in \mathcal{H}\} \geq 3$, then*

$$\mathcal{E}_k(\mathcal{H}) = \left(1 + \frac{1}{\rho - 2} + o(1)\right) k.$$

This theorem will follow as a corollary of Theorem 4.3.5.

Recalling our motivation from the Zarankiewicz problem, we show that complete bipartite graphs are optimal hosts for at least one natural family, namely the collection $\mathcal{C}_e := \{C_4, C_6, \dots\}$ of even cycles:

Theorem (See Theorem 4.2.10)**.** *For $k \geq 4$, $\mathcal{E}_k(\mathcal{C}_e) = \left\lfloor \frac{k^2}{4} \right\rfloor$, with $K_{\lfloor k/2 \rfloor, \lceil k/2 \rceil}$ being the unique extremal graph for $k \geq 6$.*

On the other hand, this is already a challenge for the case $H = K_{2,2}$:

Question 4.1.2. *What is $\mathcal{E}_k(C_4)$ and what is the optimal host graph?*

One peculiar feature of our question is that it is sensible even for *multigraphs* (graphs with potentially more than one edge between vertices) or *nonuniform hypergraphs* (where edges need not contain the same number of vertices). We let $\mathcal{E}_k^*(\mathcal{H})$ denote the maximum number of edges among host multigraphs G with $\mathrm{ex}(G, \mathcal{H}) < k$. The parameter $\mathcal{E}_k^*(\mathcal{H})$ will be important in proving bounds on $\mathcal{E}_k(\mathcal{H})$ when $\mathcal{H}$ is a family of simple graphs. However, we do not even know the following:

Conjecture (See Section 4.2.3)**.** *If $\mathcal{H}$ consists only of simple graphs, then $\mathcal{E}_k(\mathcal{H}) = \mathcal{E}_k^*(\mathcal{H})$.*

Curiously, for *non-uniform* graphs H without parallel edges, the above conjecture fails:

Theorem (See Theorems 4.3.10 & 4.3.12)**.** *Let $\mathcal{O}_2$ be the graph with a single edge and a loop at each end. Then $\mathcal{E}_k(\mathcal{O}_2) = \frac{3k}{2}$, whereas $\mathcal{E}_k^*(\mathcal{O}_2) \sim \phi k$, where ϕ is the golden ratio.*

In our study of $\mathcal{E}_k(\mathcal{H})$ and optimal host graphs, we will also show that:

1. Cliques are best at forcing cliques (Theorem 4.2.6),
2. Cliques are best at forcing a cycle (Theorem 4.2.8),
3. Complements of matchings are best at forcing $\{P_3, K_3\}$ (Theorem 4.2.12),
4. Cliques with pendant edges are best at forcing P_3 (Theorem 4.2.18),
5. Two disjoint cliques or a modified power of a cycle, depending on parity, are best at forcing $P_1 \cup P_2$ (Corollary 4.2.21 & Theorem 4.2.22),
6. For uniform hypergraphs H, $\mathcal{E}_k(H)$ is only infinite for sunflowers (Proposition 4.3.2),
7. For 1-uniform "multigraphs" H, $\mathcal{E}_k^*(H)$ is quadratic in k (Theorem 4.3.15).

In fact, for items 1, 3, and 5, the correct behavior of $\mathcal{E}_k(\mathcal{H})$ is implicit in references [5], [24] and [1], respectively, but our results will prove uniqueness of the respective host graphs.

The organization of this chapter is as follows. In Section 4.2, we begin our study of $\mathcal{E}_k(\mathcal{H})$ by obtaining the natural analogue of Turán's theorem. We then explore $\mathcal{E}_k(\mathcal{H})$ when $\mathcal{H}$ is a family of cycles and when $\mathcal{H}$ consists of small graphs, in some cases extending the results to $\mathcal{E}_k^*(\mathcal{H})$. In Section 4.3, we then explore $\mathcal{E}_k(\mathcal{H})$ when $\mathcal{H}$ is a family of hypergraphs. In addition to uniform hypergraphs, Section 4.3 also considers our problem in the context of non-uniform hypergraphs and 1-uniform multigraphs. Finally, in Section 4.4, we present conjectures and future directions.

4.1.1 Notation

We follow standard notation from [45]. For a graph $G = (V, E)$ and $S, T \subseteq V$, we use $G[S]$ to denote the subgraph of G induced by S and $G[S, T]$ to denote the subgraph of G with vertex set $S \cup T$ where $xy \in E(G[S, T])$ if and only if $xy \in E$ and $x \in S$ and $y \in T$. For a graph G and integer t, we denote the graph consisting of t vertex-disjoint copies of G by tG, e.g. tK_2 is the matching on t edges. For integers $m \leq n$, we use $[m, n] = \{m, m+1, \ldots, n\}$ and $[n] = [1, n]$. In contrast to [45], P_t will denote the path on t *edges*. Additionally, unless stated otherwise, all graphs throughout this chapter will be assumed to have no isolated vertices.

4.2 Graphs and Multigraphs

A natural starting point with the study of Turán-type questions is to consider equivalent versions of the theorems of Turán [43] and Erdős-Stone [20]. Theorem 4.1.1 follows very easily from Erdős-Stone, but we can show a much broader result in the setting of hypergraphs (see Theorem 4.3.5), and so the proof is postponed until Section 4.3. As such, we begin our study of the parameter $\mathcal{E}_k(\mathcal{H})$ with $\mathcal{H} = \{K_t\}$ as per Turán, where we can also classify the extremal graphs.

In order to do so, we establish two new definitions and a lemma which will also be used in subsequent results. Although the Turán problem is uninteresting when H is a multigraph, the parameter $\mathcal{E}_k(\mathcal{H})$ leads to fruitful questions. To this end, if $\mathcal{H}$ is a family of (multi)graphs, define

$$\mathcal{E}_k^*(\mathcal{H}) := \sup\{|E(G)| : G \text{ a multigraph and } \mathrm{ex}(G, \mathcal{H}) < k\}.$$

If $\mathcal{H}$ consists only of simple graphs, it is easy to observe that $\mathcal{E}_k(\mathcal{H}) \leq \mathcal{E}_k^*(\mathcal{H})$, so we can often consider the latter parameter instead. It is unclear whether $\mathcal{E}_k(\mathcal{H}) = \mathcal{E}_k^*(\mathcal{H})$ for every family of simple graphs $\mathcal{H}$, and we will discuss this further in Section 4.2.3

Definition 4.2.1. If G is a multigraph and $I \subseteq V(G)$, define $G' = C_I(G)$ to be the multigraph with the same number of edges obtained by contracting together the vertices in I. More specifically, write $V(G') := (V(G) \cup \{z\})\backslash I$ for some new vertex z, and the multiset $E(G') := \{C_I(e) : e \in E(G)\}$, where

$$C_I(e) := \begin{cases} zz & \text{if } e \in \binom{I}{2}; \\ zx & \text{if } e = ux \text{ for some } u \in I; \\ e & \text{otherwise.} \end{cases}$$

Here, we think of C_I as a bijection between multigraph edge sets.

To apply contractions in determining $\mathcal{E}_k^*(\mathcal{H})$, we provide the following general definition and lemma.

Definition 4.2.2. If $\mathcal{G}$ denotes the space of all finite simple graphs and $\mathcal{G}^*$ denotes the space of all finite multigraphs, a function $f : \mathcal{G}^* \to \mathcal{G}$ is called a *graph simplification* if it preserves vertex sets and containment. That is, for every pair of graphs G, H, $V(f(G)) = V(G)$ and if $H \subseteq G$, then $f(H) \subseteq f(G)$.

Examples include:

1. $f(G) = G_s$ where G_s is the underlying simple graph of G.
2. $ab \in E(f(G)) \Leftrightarrow a, b$ in the same connected component of G,
3. $ab \in E(f(G)) \Leftrightarrow \mathrm{dist}_G(a, b) \leq t$ for some fixed integer t.

Lemma 4.2.3. *Let f be a multigraph simplification such that $f(H)$ is a clique for every $H \in \mathcal{H}$. By contrast, let G be a graph and I be an independent set in $f(G)$. If $G' = C_I(G)$, then $\mathrm{ex}(G', \mathcal{H}) \leq \mathrm{ex}(G, \mathcal{H})$.*

Note that an independent set in $f(H)$ is not necessarily an independent set in H, as seen by e.g. $u \sim v$ in $f(G) \Leftrightarrow u, v$ are 2-connected in G. However, all of the scenarios in which we will use this lemma (namely, the three examples given above), $f(H) \supseteq H$ for every H. In particular, we will never be contracting edges in G to loops, as per the first case in the definition of $C_e(I)$ above.

Proof. It suffices to show that if some $F \subseteq G$ contains a copy of $H \in \mathcal{H}$, then $C_I(F) \subseteq G'$ still contains a copy of some $H' \in \mathcal{H}$. In fact, more is true; namely, if $H_0 \subseteq G$ is a copy of H, then $C_I(H_0) \subseteq G'$ contains a copy of H. To see this, as f is a graph simplification, $f(H) \simeq f(H_0) \subseteq f(G)$, so as $f(H)$ is a clique, $|I \cap V(H_0)| \leq 1$. In other words, $C_I(H_0)$ is a copy of H, possibly with extra multiedges or loops. □

For a graph simplification f, we say that G is *f-compressed* if $f(G)$ is a clique. Further, we say that G is an *f-compressed copy of G'* if G is f-compressed and there is a sequence of graphs $G' = G_0, G_1, \ldots, G_t = G$ such that $G_{i+1} = C_I(G_i)$ for some independent set I in $f(G_i)$. Note that if G is an f-compressed copy of G', then $|E(G)| = |E(G')|$. With this definition, the following corollary follows immediately from Lemma 4.2.3.

Corollary 4.2.4. *Suppose, as above, that f is a multigraph simplification where $f(H)$ is a clique for every $H \in \mathcal{H}$. If G^* is an f-compressed copy of G, then $\mathrm{ex}(G^*, \mathcal{H}) \leq \mathrm{ex}(G, \mathcal{H})$. In particular, when computing $\mathcal{E}_k^*(\mathcal{H})$, it suffices to consider graphs G such that $f(G)$ is a clique, i.e. $\mathcal{E}_k^*(\mathcal{H}) = \sup\{|E(G)| : \mathrm{ex}(G, \mathcal{H}) < k, f(G) \simeq K_{|V(G)|}\}$.*

Before finding the value of $\mathcal{E}_k(K_t)$, we first must recall some properties of Turán graphs. Define $T_{t-1}(n)$ to be the balanced complete $(t-1)$-partite graph on n vertices; Turán's Theorem states that $\mathrm{ex}(K_n, K_t) = |E(T_{t-1}(n))|$. Additionally, define the *Turán density* of K_t in K_n by $\alpha_n(t) := \mathrm{ex}(K_n, K_t)/\binom{n}{2}$. We will use the following observations in the subsequent proof.

Observation 4.2.5. *If $n \equiv n_0 \pmod{t-1}$, then*

$$|E(T_{t-1}(n))| = \binom{n}{2} - n_0\binom{\frac{n-n_0}{t-1}+1}{2} - (t-1-n_0)\binom{\frac{n-n_0}{t-1}}{2} = \left(1 - \frac{1}{t-1} \pm O\left(\frac{1}{n}\right)\right)\binom{n}{2}.$$

As such, if $(t-1) \nmid n$, we have

$$|E(T_{t-1}(n))| = |E(T_{t-1}(n-1))| + (n-1) - \left\lfloor \frac{n-1}{t-1} \right\rfloor.$$

In particular, this implies that if $(t-1) \nmid n$, then $\alpha_{n-1}(t) > \alpha_n(t)$. Furthermore, $\alpha_{n-1}(t) \geq \alpha_n(t)$ for all n, which can be seen by averaging over subgraphs.

The following proof uses an idea by Alon (see [5, Lemma 2.1]) in the context of chromatic numbers.

Theorem 4.2.6. *For any integer $t \geq 3$,*

$$\mathcal{E}_k(K_t) = \mathcal{E}_k^*(K_t) = \left(1 + \frac{1}{t-2} + o(1)\right)k.$$

Moreover, for infinitely many values of k, the unique extremal graph for $\mathcal{E}_k^*(K_t)$ and $\mathcal{E}_k(K_t)$ is a clique.

Proof. Lower bound. For any positive integer k, let n be the largest integer for which $k > \mathrm{ex}(K_n, K_t)$. As $\mathrm{ex}(K_n, K_t) = \left(1 - \frac{1}{t-1} \pm O\left(\frac{1}{n}\right)\right)\binom{n}{2}$, we observe that $\mathrm{ex}(K_{n+1}, K_t) - \mathrm{ex}(K_n, K_t) = O(n)$. Thus, $k \leq \mathrm{ex}(K_n, K_t) + O(n) = \mathrm{ex}(K_n, K_t) + O(\sqrt{k})$, so we calculate

$$\mathcal{E}_k(K_t) \geq \binom{n}{2} \geq \frac{k - O(\sqrt{k})}{\mathrm{ex}(K_n, K_t)}\binom{n}{2} = \left(1 + \frac{1}{t-2} + o(1)\right)k.$$

Upper bound. Let G be a (multi)graph with $\mathrm{ex}(G, K_t) < k$. Letting f be the "underlying simple graph" simplification ((1) in Definition 4.2.4), as $f(K_t) = K_t$, we may suppose that G is f-compressed by Corollary 4.2.4. In other words, G is a clique, possibly with parallel edges. Let $n = |V(G)|$ and write $|E(G)| = \binom{r}{2} + \ell$ where $0 \leq \ell \leq r - 1$. As G is a copy of K_n, possibly with parallel edges, we know that $r \geq n$.

Now, let T be a copy of the Turán graph $T_{t-1}(n)$ chosen uniformly at random on $V(G)$, and let H be the multigraph with edge set $\{uv \in E(G) : uv \in E(T)\}$ (so that if u, v span multiple edges in G then they either all survive the intersection with T or all do not). As any such H is K_t-free, writing $\alpha_n = \alpha_n(t)$, we calculate

$$\mathrm{ex}(G, K_t) \geq \mathbf{E}|E(H)| = |E(G)| \cdot \alpha_n \geq |E(G)| \cdot \alpha_r = \left(\binom{r}{2} + \ell\right) \cdot \frac{|E(T_{t-1}(r))|}{\binom{r}{2}} = \mathrm{ex}(K_r, K_t) + \ell\alpha_r. \tag{4.1}$$

Thus, for any positive integer k, let r be the least integer for which $k \leq \mathrm{ex}(K_r, K_t) + 1$. As above, we note that $k \geq \mathrm{ex}(K_r, K_t) - O(\sqrt{k})$. Equation (4.1) shows that for any multigraph G, if $|E(G)| > \binom{r}{2}$, then $\mathrm{ex}(G, K_t) \geq k$, so

$$\mathcal{E}_k^*(K_t) \leq \binom{r}{2} \leq \frac{k + O(\sqrt{k})}{\mathrm{ex}(K_r, K_t)}\binom{r}{2} = \left(1 - \frac{1}{t-2} + o(1)\right)k.$$

As $\mathcal{E}_k(K_t) \leq \mathcal{E}_k^*(K_t)$, this establishes the asymptotics. In particular, we have shown that if $k = \mathrm{ex}(K_r, K_t) + 1$ for some integer r, then $\mathcal{E}_k(K_t) = \mathcal{E}_k^*(K_t) = \binom{r}{2}$.

Extremal graphs. We now wish to show that for infinitely many k, the only extremal graph for $\mathcal{E}_k^*(K_t)$, and thus for $\mathcal{E}_k(K_t)$, is a clique.

Let $k = \text{ex}(K_r, K_t) + 1$ where $r > t$ and $(t-1) \nmid r$. In this case, we know that $\mathcal{E}_k(K_t) = \mathcal{E}_k^*(K_t) = \binom{r}{2}$ and that $\alpha_n > \alpha_r$ for any $n < r$. Now, let G be an f-compressed graph which is extremal for $\mathcal{E}_k^*(K_t)$. As before, $|V(G)| \leq r$, and as $\alpha_n > \alpha_r$ for any $n < r$, the only way for $\text{ex}(G, K_t) \leq k - 1 = \text{ex}(K_r, K_t)$ is if $|V(G)| = r$, as shown by Equation (4.1). Thus, as G has $\binom{r}{2}$ edges, r vertices and contains K_r, it must be the case that $G \simeq K_r$.

Now, suppose G is any graph on $\binom{r}{2}$ edges with $\text{ex}(G, K_t) < k$. Let $G = G_0, G_1, \ldots, G_q = G^*$ where G^* is f-compressed and $G_{i+1} = C_{xy}(G_i)$ for some $xy \notin E(G_i)$. By the above argument, we know that $G^* \simeq K_r$. Now, suppose $G \not\simeq K_r$; so that $q \geq 1$. Let $u, v \in V(G_{q-1})$ be such that $G^* = C_{uv}(G_{q-1})$. For ease of notation, we will write $N(x) = N_{G_{q-1}}(x)$ for the remainder of the proof.

As G_{q-1} can be contracted once more, $G_{q-1} \not\simeq K_r$. Then $|V(G_{q-1})| > r$ and as $G^* = K_r$, we must have $N(u) \cup N(v) = V(G_{q-1}) \setminus \{u, v\}$ and $V(G_{q-1}) \setminus \{u, v\}$ must induce a copy of K_{r-1}. Further, G^* is simple, so it must be the case that G_{q-1} is simple, moreover $N(u) \cap N(v) = \varnothing$ otherwise G^* would contain a multiedge upon contracting uv. We check that such a graph has a K_t-free subgraph which is too large.

Indeed, first suppose $|N(u)| < \lfloor \frac{r-1}{t-1} \rfloor$. Then let T be a copy of $T_{t-1}(r-1)$ contained in $V(G_{q-1}) \setminus \{u, v\}$ with parts $X_1, \ldots, X_{t-1}$ where $X_1 \supseteq N(u)$ and $|X_1| = \lfloor \frac{r-1}{t-1} \rfloor$. Then if H is the subgraph consisting of the edges in T along with the edges incident to u and edges of the form $\{vx : x \notin X_{t-1}\}$, we find that $H \subseteq T_{t-1}(r+1)$ as $uv \notin E(G_{q-1})$, so H is K_t-free. Additionally,

$$\begin{aligned} |E(H)| &= |E(T_{r-1}(r-1))| + |N(u)| + |N(v)| - |X_1 \cap N(v)| \\ &\geq |E(T_{t-1}(r-1))| + (r-1) - \left(\left\lfloor \frac{r-1}{t-1} \right\rfloor - 1 \right) \\ &= |E(T_{t-1}(r))| + 1 = k, \end{aligned}$$

a contradiction. Thus, we may suppose that $|N(u)|, |N(v)| \geq \lfloor \frac{r-1}{t-1} \rfloor$. Additionally, as $|N(u)| + |N(v)| = r - 1$, we have, without loss of generality, $|N(v)| \geq \lceil \frac{r-1}{t-1} \rceil$. As such, let T be a copy of $T_{t-1}(r-1)$ contained in $V(G_t) \setminus \{u, v\}$ with parts $X_1, \ldots, X_{t-1}$ where $X_1 \subseteq N(u)$ and $X_2 \subseteq N(v)$. Now, let H consist of T along with all edges incident to u or v. As $uv \notin E(G_{q-1})$, H is again a subgraph of $T_{t-1}(r+1)$, and so is K_t-free. However,

$$\begin{aligned} |E(H)| &= |E(T_{t-1}(r-1))| + |N(u)| + |N(v)| \\ &= |E(T_{t-1}(r-1))| + r - 1 \\ &= |E(T_{t-1}(r))| + \left\lfloor \frac{r-1}{t-1} \right\rfloor \geq k, \end{aligned}$$

another contradiction. We conclude that any (multi)graph G with $|E(G)| = \binom{r}{2}$ and $\text{ex}(G, K_t) < k$ must be a copy of K_r. □

It is not clear what the precise value of $\mathcal{E}_k(K_t)$ and $\mathcal{E}_k^*(K_t)$ are when $k \neq \text{ex}(K_r, K_t) + 1$ for any r, but we conjecture the following:

Conjecture 4.2.7. *For positive integers $r_1 \geq \cdots \geq r_\ell$, let $K(r_1, \ldots, r_\ell)$ be the multigraph consisting of "nested" copies of K_{r_i}: that is, on vertex set $[r_1]$, we overlay a copy of K_{r_i} on $[r_i]$ for every i (thus, the maximum edge-weight is ℓ, provided every $r_i \geq 2$). For every k, there exist positive integers $r_1 \geq \cdots \geq r_\ell$ such that $K(r_1, \ldots, r_\ell)$ is extremal for $\mathcal{E}_k^*(K_t)$.*

4.2.1 Cycles

We begin this section with a simple result related to the graphic matroid rank of a graph.

Theorem 4.2.8. *If $\mathcal{C} := \{C_3, C_4, \ldots\}$ is the set of all cycles, then $\mathcal{E}_k(\mathcal{C}) = \binom{k}{2}$. Furthermore, the only extremal graph for $\mathcal{E}_k(\mathcal{C})$ is K_k.*

Proof. Note that any k-edge subgraph of K_k contains a cycle, hence $\mathcal{E}_k(\mathcal{C}) \geq \binom{k}{2}$. Now, suppose that G is some connected graph with $|E(G)| > \binom{k}{2}$, then $E(G)$ spans at least $k+1$ vertices. As such, G has a spanning tree with at least k edges, so $\mathrm{ex}(G, \mathcal{C}) \geq k$. Hence, any connected G with $\mathrm{ex}(G, \mathcal{C}) < k$ has $|E(G)| \leq \binom{k}{2}$. Since every cycle is connected, we are done by taking f to be the connectedness simplification ((2) in Definition 4.2.2) in Corollary 4.2.4.

We now wish to argue that the only extremal graph for $\mathcal{E}_k(\mathcal{C})$ is K_k. The above shows the only connected G with $\binom{k}{2}$ edges and $\mathrm{ex}(G, \mathcal{C}) < k$ is K_k. On the other hand, suppose there were some disconnected G with $|E(G)| = \binom{k}{2}$ and $\mathrm{ex}(G, \mathcal{C}) < k$; then fixing some $I \subseteq V(G)$ with one vertex in each connected component gives $C_I(G) \simeq K_k$ by this uniqueness and Corollary 4.2.4. However, this implies that K_k has a cut-vertex, which is not true. Thus, G must have been connected in the first place, so $G \simeq K_k$. □

In fact, we know that multigraphs are no better at forcing cycles:

Corollary 4.2.9. *If a multigraph G has $\mathrm{ex}(G, \mathcal{C}) < k$, then $|E(G)| \leq \binom{k}{2}$, with equality if and only if $G \simeq K_k$. In particular, $\mathcal{E}_k^*(\mathcal{C}) = \binom{k}{2}$.*

Proof. By the same logic as before, we may begin by assuming G is connected.

Again, any (simple) spanning tree in G is $\mathcal{C}$-free, so $\mathrm{ex}(G, \mathcal{C}) < k \Rightarrow |V(G)| \leq k$. Furthermore, the set of edges incident to any fixed v is also $\mathcal{C}$-free, so $\Delta(G) \leq k-1$. Thus

$$|E(G)| = \frac{1}{2}\sum_{v \in V} \deg(v) \leq \frac{1}{2}|V(G)|\Delta(G) \leq \frac{1}{2}k(k-1).$$

If we have equality, then certainly $|V(G)| = k$. But there cannot be any edge of multiplicity 2 or higher; otherwise, extending this to a spanning tree of G (with multi-edges) will gain a further $k-2$ edges at least, yielding a $\mathcal{C}$-free subgraph of G with at least k edges. Thus G was simple, and hence must be K_k by Theorem 4.2.8. □

If $\mathcal{H}$ does not contain a bipartite graph, then the asymptotic value of $\mathcal{E}_k(\mathcal{H})$ is determined by Theorem 4.1.1, which will be proved in a more general context later (see Theorem 4.3.5). Thus, it is natural about this extremal function for the class of all even cycles, denoted by $\mathcal{C}_e$.

Theorem 4.2.10. *For $k \geq 4$, $\mathcal{E}_k(\mathcal{C}_e) = \left\lfloor \frac{k^2}{4} \right\rfloor$. Furthermore, the only extremal graph for $\mathcal{E}_k(\mathcal{C}_e)$ is the balanced complete bipartite graph on k vertices, unless $k = 5$.*

Proof. Lower bound. Let G be the balanced complete bipartite graph on k vertices. Naturally, any k edges from G contain a cycle, which is necessarily even as G is bipartite. Hence, $\mathcal{E}_k(\mathcal{C}_e) \geq |E(G)| = \lfloor \frac{k}{2} \rfloor \lceil \frac{k}{2} \rceil = \lfloor \frac{k^2}{4} \rfloor$.

For the upper bound, we again look to use Corollary 4.2.4, and first prove the connected case.

Lemma 4.2.11. *If G is connected with $|E(G)| \geq \left\lfloor \frac{k^2}{4} \right\rfloor$, then $\mathrm{ex}(G, \mathcal{C}_e) \geq k-1$, with equality if and only if $G = K_{\lceil k/2 \rceil, \lfloor k/2 \rfloor}$ or, in the case of $k = 5$, $G = K_4$.*

Proof. Let G be any connected graph on n vertices with $\mathrm{ex}(G, \mathcal{C}_e) \leq k-1$. For any spanning tree F of G, F contains no even cycle, so $|E(F)| \leq k-1$, or in other words, $n \leq k$. As such, set $k = n + q$, and assume

$$|E(G)| \geq \left\lfloor \frac{k^2}{4} \right\rfloor = \frac{n^2}{4} + \frac{2nq + q^2 - \mathbf{1}_{\text{k odd}}}{4} \geq \frac{n^2}{4} + \frac{2nq + q^2 - 1}{4},$$

but that G is not the complete balanced bipartite graph. Then as $n \leq k$, we know, by the uniqueness of the Turán graph, that G contains a triangle. We will attempt to use the triangles in G to build a large $\mathcal{C}_e$-free subgraph.

Say $T \subseteq G$ is a "triangle forest" with t triangles if $E(T)$ is a collection of t edge-disjoint triangles such that the removal of any one edge from each triangle forms a forest. In particular, the only cycles within such a T are the t triangles. So we may extend T to a spanning subgraph H (using connectivity) with no additional cycles, thus H is still $\mathcal{C}_e$-free. We deduce that $(n-1) + t = |E(H)| \leq k-1$, so we must have $t \leq q$. In particular, if $q = 0$, then G must be the balanced complete bipartite graph on k vertices. Thus, for the remainder of the proof, we shall suppose $q \geq 1$.

Now, take such a triangle forest T with:

1. $|E(T)|$ (and hence t) as large as possible,
2. Subject to (1), if $T = T_1 \cup \cdots \cup T_\ell$ is a decomposition of T into connected components where $|T_1| \geq \cdots \geq |T_\ell|$, then $(|T_1|, \ldots, |T_\ell|)$ is maximal in the lexicographic ordering.

By the lexicographic order, we mean that $(a_1, \ldots, a_\ell) \succ (b_1, \ldots, b_{\ell'}) \Leftrightarrow a_j > b_j$ for $j := \min\{i : a_i \neq b_i\}$. Such a lexicographic maximal T means there is no $v \in T_i$ with 2 edges to

the same triangle in T_j for any $i < j$. If this were not the case and wxy was the triangle with both $vx, vy \in E(G)$, then let $T' := (T \cup \{vx, vy\}) \backslash \{wx, wy\}$ (see Figure 4.1). T' is a triangle forest with the same number of edges as T, with $|T'_j| = |T_j|$ for all $j < i$ yet $|T'_i| \geq |T_i| + 2$, so T' is lexicographically larger than T, contradicting (2).

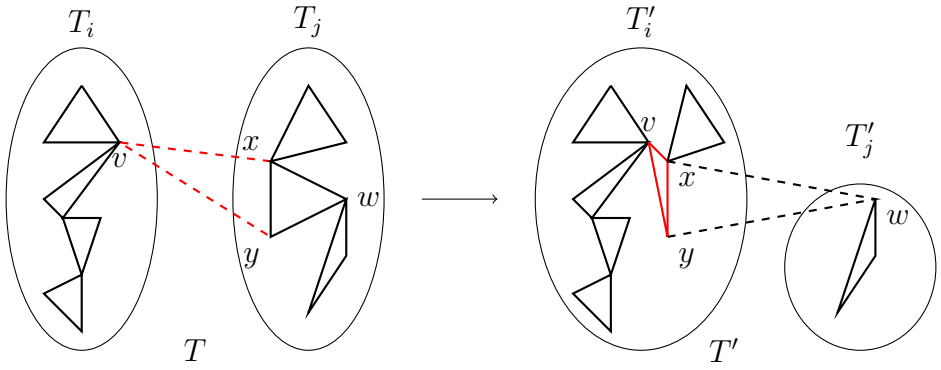


Figure 4.1: Finding a lexicographically larger triangle forest in the case where some vertex in T_i has two edges to the same triangle in T_j.

Thus, if T_j consists of t_j triangles for every j (so that $|T_j| = 2t_j + 1$ and $t = \sum_j t_j$), then whenever $i < j$, every $v \in T_i$ has at most t_j edges to T_j. Summing over all $v \in T_i$ gives $|E[T_i, T_j]| \leq (2t_i + 1)t_j$.

We now attempt to bound the remaining edges in G. Crudely, $|E(G[T_i])| \leq \binom{|T_i|}{2} = 2t_i^2 + t_i$ for all i.

Case 1: $|V(T)| \leq \frac{n}{2}$.
Let $G' := G \setminus \bigcup_i G[T_i]$. As T is maximal, G' must be triangle-free, so certainly $|E(G')| \leq \frac{n^2}{4}$. Therefore,

$$\frac{n^2}{4} + \frac{2nq + q^2 - 1}{4} \leq |E(G)| = |E(T)| + |E(G')| \leq \sum_{i=1}^{\ell} (2t_i^2 + t_i) + \frac{n^2}{4},$$

and so

$$\frac{2nq + q^2 - 1}{4} \leq t_1 \sum_{i=1}^{\ell} (2t_i + 1) = t_1 |V(T)| \leq t|V(T)| \leq q \cdot \frac{n}{2}.$$

Thus, $q = 1$ as we supposed that $q \geq 1$, so we have equality everywhere. In particular, $t_1 = t = q = 1$, so T is a single triangle, $|V(T)| = \frac{n}{2} \Rightarrow n = 6$, and $|E(G')| = \frac{n^2}{4} \Rightarrow G' = K_{3,3}$. Since G is therefore a 6-vertex, edge-disjoint union of $K_{3,3}$ with a triangle, this uniquely determines G as $K_6 \backslash K_3$, and this G still has $\mathrm{ex}(K_6 \backslash K_3, \mathcal{C}_e) \geq 7 = n + q = k$ (see Figure 4.2); a contradiction.

Case 2: $|V(T)| > \frac{n}{2}$.
In this case, for the triangle-free graph $G'' := G \backslash G[T] = G' \backslash \bigcup_{i,j} G[T_i, T_j]$, $V(T)$ spans an independent set in G'', so we can apply a stronger version of the Mantel bound (reproved

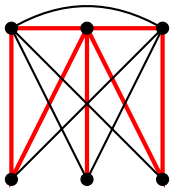


Figure 4.2: A $\mathcal{C}_e$-free subgraph of $K_6 \backslash K_3$ with 7 edges.

here for completeness): for each $v \in V$, $\deg_{G''}(v) \leq \alpha(G'') = \alpha$. Now, if I is an independent set of size α, then every edge of G'' must meet $V \setminus I$, so

$$|E(G'')| \leq \sum_{v \in V \setminus I} \deg_{G''}(v) \leq |I||V \setminus I| = \alpha(n - \alpha).$$

Of course, $V(T)$ is an independent set in G'' by construction, so $\alpha \geq |V(T)| = 2t + \ell$. As $x(n-x)$ is strictly decreasing for $x \geq n/2$, we have $|E(G'')| \leq (2t+\ell)(n-(2t+\ell))$ as $2t + \ell > \frac{n}{2}$.

We run a similar calculation in this case:

$$\begin{aligned}
\left\lfloor \frac{k^2}{4} \right\rfloor \leq |E(G)| &\leq \sum_{i=1}^{\ell} |E(G[T_i])| + \sum_{i<j} |E(T_i, T_j)| + |E(G'')| \\
&\leq \sum_{i=1}^{\ell} \left(2t_i^2 + t_i\right) + \sum_{i<j} \left((2t_i+1)t_j\right) + \left(n - (2t+\ell)\right)(2t+\ell) \\
&\leq \sum_{i=1}^{\ell} (2t_i^2 + t_i) + \sum_{i \neq j} \left(t_i t_j + \frac{t_i + t_j}{4} \right) + n(2t+\ell) - (2t+\ell)^2 \\
&= t^2 + \sum_{i=1}^{\ell} t_i^2 + \left(\frac{\ell+1}{2} \right) t + n(2t+\ell) - (2t+\ell)^2,
\end{aligned}$$

so

$$\begin{aligned}
&n^2 + 2qn + q^2 - 4n(2t+\ell) + 4(2t+\ell)^2 - \mathbf{1}_{k \text{ odd}} \leq 8t^2 + 2(\ell+1)t \\
&\quad \Rightarrow \left(n + q - 2(2t+\ell)\right)^2 + 4q(2t+\ell) - \mathbf{1}_{k \text{ odd}} \leq 8t^2 + 2(\ell+1)t \leq 8qt + 2(2\ell)q.
\end{aligned}$$

It follows $|k - 2(2t+\ell)| \leq \mathbf{1}_{k \text{ odd}}$. But the reverse is true whether k is even or odd, hence we again obtain all inequalities above at equality. So certainly $t = q$, $\ell = 1$, $G[V(T)] = G[V(T_1)]$ is a clique, and $\alpha(G'') = 2t + \ell$, so $|E(G'')| = \sum_{v \notin V(T)} \deg_{G''}(v) = (n - (2t+\ell))(2t+\ell)$. As such, $G''[\overline{V(T)}]$ is empty and $\deg_{G''}(v) = 2t + \ell$ for every $v \notin V(T)$, so G'' is the complete bipartite graph on $[V(T), \overline{V(T)}]$. Putting this together with the clique on $V(T)$, deduce $G \simeq K_n \backslash K_r$, where $r = n - (2t+\ell) = n - (2q+1)$.

We know $k-(4q+2) =: \epsilon \in \{0, \pm 1\}$, so $r = (k-q)-(2q+1) = q+1+\epsilon$. Now, if $r \geq q+1$, we can find a triangle forest F with $q+1$ triangles (contradicting maximality of T as $t \leq q$) by taking a path on $2(q+1)$ edges with $q+1$ vertices in $\overline{V(T)}$ and $q+2 \leq 2q+1$ among $V(T)$, and completing the $q+1$ edge-disjoint copies of P_3 into K_3's using the edges from inside T (See Figure 4.3). Furthermore, if $q \geq 2$, then we may similarly choose P by instead taking $q \leq q+1+\epsilon$ vertices of $\overline{V(T)}$ and $q+3 \leq 2q+1$ vertices of $V(T)$. Otherwise, $\epsilon = -1$ and $q = 1$. In this case, we deduce that $G \simeq K_4$, which does have $\mathrm{ex}(K_4, \mathcal{C}_e) = \mathrm{ex}(K_{2,3}, \mathcal{C}_e) = 4$. □

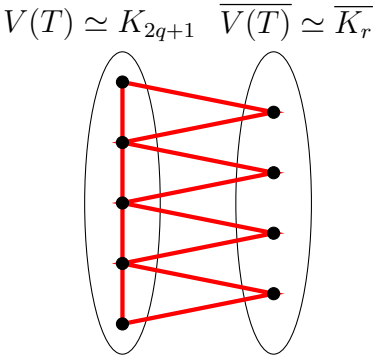


Figure 4.3: A large triangle forest contained in $K_n \setminus K_r$.

Upper bound. If G is now arbitrary with $|E(G)| \geq \left\lfloor \frac{k^2}{4} \right\rfloor$, then forming any $I \subseteq V(G)$ with one vertex from each connected component gives $\mathrm{ex}(G, \mathcal{C}_e) \geq \mathrm{ex}(C_I(G), \mathcal{C}_e) \geq k-1$. If we have equality here, we know $C_I(G)$ is necessarily $K_{\lfloor k/2 \rfloor, \lceil k/2 \rceil}$ (or K_4) by the lemma, yet none of these graphs have a cut-vertex for $k \geq 4$. Hence, G must have been connected in the first place, so G is one of the claimed extremal graphs. □

Unfortunately, the above argument is very specific to simple graphs, so we have been unable to determine $\mathcal{E}_k^*(\mathcal{C}_e)$ unless it happens to be the case that $\mathcal{E}_k^*(\mathcal{C}_e) = \mathcal{E}_k(\mathcal{C}_e)$.

4.2.2 Small Graphs

In this section, we will explore $\mathcal{E}_k(\mathcal{H})$ where $\mathcal{H}$ is a collection of small graphs. At the end of this section, we also give a complete classification of the families which have $\mathcal{E}_k(\mathcal{H}) = \infty$. Throughout this section, we will only focus on simple host graphs.

Recall that P_t denotes the path on t *edges*.

Theorem 4.2.12. *For $\mathcal{H} = \{P_3, K_3\}$, and $k \geq 3$,*

$$\mathcal{E}_k(\mathcal{H}) = \begin{cases} \binom{k+1}{2} - \frac{k+2}{2} & \text{if } k \text{ is even;} \\ \binom{k+1}{2} - \frac{k+1}{2} & \text{if } k \text{ is odd.} \end{cases}$$

Moreover, the only extremal graph for $\mathcal{E}_k(\mathcal{H})$ is

$$G_k := \begin{cases} K_{k+1} \backslash (\frac{k-2}{2} K_2 \cup P_2) & \text{if } k \text{ is even;} \\ K_{k+1} \backslash (\frac{k+1}{2} K_2) & \text{if } k \text{ is odd.} \end{cases}$$

Note that the first of these results has been previously noticed by Ferneyhough, Haas, Hanson and MacGillivray in [24, Corollary 2] using a bound on the domination number due to Vizing [44]. They also used the graphs G_k to provide the lower bounds. We offer a self-contained proof that also shows these extremal graphs are in fact unique.

Definition 4.2.13. Given a graph G, a *star-packing* of G is a subgraph of G which is a union of vertex-disjoint stars.

It is quick to observe that $H \subseteq G$ is $\{P_3, K_3\}$-free if and only if H is a star packing of G with possible isolated vertices.

Lemma 4.2.14. *Let G be a graph on $n+t$ vertices. If every star-packing in $\overline{G}$ has at most $n-2$ edges, then*

$$2|E(G)| \geq f(n,t) := \begin{cases} n + 2nt + t(t-1) & \text{if } n \text{ is even;} \\ n + 1 + 2nt + t(t-1) & \text{if } n \text{ is odd.} \end{cases}$$

Further, if equality holds, then $\overline{G} \simeq G_{n-1} \cup \overline{K_t}$.

Proof. If $n \leq 3$, the statement is straightforward, so assume $n \geq 4$. We first claim that for any $i \geq 1$ and $S \subseteq V$ with $|S| = i$, then S has at least $t - i + 2$ common neighbors in $V \setminus S$. If this were not the case, then there are at least $|V \setminus S| - (t - i + 1) = n - 1$ vertices in $V \setminus S$ which are not connected to some $v \in S$. Thus, we can find $n-1$ edges in $\overline{G}$ that form vertex-disjoint stars with centers in S, contradicting the fact that every star packing has at most $n-2$ edges. In particular this implies that

1. Taking $i = 1$, $\delta(G) = t + s + 1$ for some $s \geq 0$.
2. Taking $i = 2$, any two vertices have at least t common neighbors.

Now, proceed by induction on t.

When $t = 0$, we have $\delta(G) \geq 1$ by (1), so $2|E(G)| \geq n + \mathbf{1}_{n \text{ odd}}$, with equality if and only if $G \simeq \frac{n}{2} K_2$ when n is even or $G \simeq \frac{n-3}{2} K_2 \cup P_2$ when n is odd. In either case, $\overline{G} \simeq G_{n-2}$.

Otherwise, $t \geq 1$, so $\operatorname{diam}(G) \leq 2$ by (2). In this case, choose $v \in V$ with $\deg(v) = \delta(G) = t + s + 1$ for some $s \geq 0$ and define $N^2(v) := \{w \in G : \operatorname{dist}(v, w) = 2\} = V \setminus (N(v) \cup \{v\})$. As $\deg(v) = t + s + 1$, we have $|N^2(v)| = n - s - 2$. In particular, $\{v\} \times N^2(v)$ is a star with $n - s - 2$ edges in $\overline{G}$. Thus, setting $G' := G[N(v)]$ it must be the case that every star

packing in $\overline{G'}$ must have at most s edges, otherwise we could find a star packing in $\overline{G}$ with $n-1$ edges.

Set $n' = s+2$ and $t' = (s+t+1) - n' = t-1$. As $|V(G')| = n' + t'$, and every star packing in $\overline{G'}$ has at most $n'-2$ edges, by induction,

$$2|E(G')| \geq f(n',t') = n' + 2n't' + t'(t'-1) + \mathbf{1}_{n' \text{ odd}} = 2st - s + t^2 + t + \mathbf{1}_{\text{s odd}}.$$

Additionally, we find that $|E(G[N(v), N^2(v)])| \geq t(n-s-2)$ as $|N^2(v)| = n-s-2$ and any two vertices have at least t common neighbors. Writing $E(N(v), N^2(v))$ for the edges in the bipartite subgraph of G induced by the classes $N(v)$ and $N^2(v)$, we thus obtain

$$\begin{aligned} 2|E(N(v),N^2(v))| + 2|E(G[N^2(v)])| &= |E(N(v),N^2(v))| + \sum_{w \in N^2(v)} \deg(w) \\ &\geq \mathbf{1}_{n \text{ odd, } s \text{ even}} + t(n-s-2) + (n-s-2)(t+s+1) \\ &= \mathbf{1}_{n \text{ odd, } s \text{ even}} + (n-s-2)(2t+s+1), \end{aligned} \tag{4.2}$$

since $(n-s-2)(2t+s+1)$ is odd whenever both n is odd and s is even. So, we calculate

$$\begin{aligned} 2|E(G)| &= 2|E(N(v),N^2(v))| + 2|E(G[N^2(v)])| + 2\deg(v) + 2|E(G')| \\ &\geq \mathbf{1}_{n \text{ odd, } s \text{ even}} + (n-s-2)(2t+s+1) + 2(t+s+1) + f(n',t') \\ &\geq \mathbf{1}_{n \text{ odd, } s \text{ even}} + (n-s-2)s + \big((n-s-2)(2t+1) + 2(t+s+1)\big) + \\ &\qquad \big(2st - s + t^2 + t + \mathbf{1}_{s \text{ odd}}\big) \\ &= \mathbf{1}_{n \text{ odd, } s \text{ even}} - \mathbf{1}_{n \text{ odd}} + \mathbf{1}_{s \text{ odd}} + (n-s-2)s + \big(n + 2nt + t(t-1) + \mathbf{1}_{n \text{ odd}}\big) \\ &= \mathbf{1}_{n \text{ odd, } s \text{ even}} - \mathbf{1}_{n \text{ odd}} + \mathbf{1}_{s \text{ odd}} + (n-s-2)s + f(n,t) \\ &\geq f(n,t). \end{aligned}$$

The last inequality follows from $n - s - 2 = |N^2(v)| \geq 0$. We have now established the claimed bound.

When $2|E(G)| = f(n,t)$, we wish to show $\overline{G} \simeq G_{n-1} \cup \overline{K_t}$. Certainly, all inequalities above are equalities, so $s|N^2(v)| = 0$. If $N^2(v) = \varnothing$, then $\delta(G) = \deg(v) = |V| - 1$; hence, $G \simeq K_{n+t}$, a contradiction as $2|E(K_{n+t})| > f(n,t)$ whenever $n \geq 2$.

So instead $s = 0$. Deduce $\delta(G) = \deg(v) = t+1$ by (1), and for any $w \in V$, $|N(v) \cap N(w)| \geq t$ by (2). As such, $G' = G[N(v)] \simeq K_{t+1}$.

Equality in Equation (4.2), shows that all but (at most) one $w \in N^2(v)$ satisfy both $d(w) = t+1$ and $|N(w) \cap N(v)| = t$, and thus has exactly 1 edge inside $N^2(v)$. There are at least $|N^2(v)| - 1 = n - 3 \geq 1$ such w, so fix one such w_1 and let w_0 be its unique neighbor in $N^2(v)$. Then, w_0, w_1 share t neighbors, which must therefore be some $S \subset N(v)$.

Case 1. $N(w_0) = S \cup \{w_1\}$ (i.e. $\deg(w_0) = t+1$). Then every other $w \in N^2(v) \backslash \{w_0, w_1\}$ shares t neighbors with w_0, none of which are w_1, so must share S.

Case 2. $\deg(w_0) > t+1$. So the equality in Euqation (4.2) in fact shows $\deg(w) = t+1$ and $|N(w) \cap N(v)| = t$ for every $w \in N^2(v) \backslash \{w_0, w_1\}$. If some such w did not have S as its t neighbors in $N(v)$, then since w_2 shares t neighbors with both w_1 and w_0, it must be adjacent to both w_0 and some $w' \in N^2(v) \cap N(w_0)$ (possibly w_1). So in total, $\deg(w) \geq t+2$; a contradiction.

In either case, every vertex in S is connected to every vertex in G, so S is a collection of isolated vertices in $\overline{G}$. As such, $\overline{G} \backslash S$ still has no star-packing with at least $n-2$ edges, while $G \backslash S$ is left with $\frac{f(n,t)}{2} - \binom{t}{2} - nt = \left\lceil \frac{n}{2} \right\rceil$ edges. Crudely $\Delta(\overline{G} \backslash S) \leq n-2$, so $\delta(G \backslash S) \geq 1$, hence $G \backslash S \simeq \frac{n}{2} K_2$ (or $\frac{n-3}{2} K_2 \cup P_2$ if n is odd). Adding S back shows $\overline{G} \simeq G_{n-1} \cup \overline{K_t}$. □

Proof of Theorem 4.2.12. Lower bound. As $\Delta(G_k) = k-1$, any single star in G_k hast at most $k-1$ edges. Additionally, as $|V(G_k)| = k+1$, any star-packing in G_k with $i \geq 2$ stars has at most $k+1-i \leq k-1$ edges. Thus, $\mathrm{ex}(G_k, \{K_3, P_3\}) < k$, so $\mathcal{E}_k(\{K_3, P_3\}) \geq |E(G_k)| = \binom{k+1}{2} - \frac{k+1+\mathbf{1}_{k \text{ even}}}{2}$.

Upper bound. Let G be a graph with $\mathrm{ex}(G, \{K_3, P_3\}) < k$. Thus, every star-packing in G has at most $k-1$ edges. If G has at most k vertices, then

$$|E(G)| \leq \binom{k}{2} < \binom{k+1}{2} - \frac{k+1+\mathbf{1}_{k \text{ even}}}{2}.$$

Thus, we may suppose G has $k+1+t$ vertices for some $t \geq 0$. By Lemma 4.2.14, if every star packing in G has at most $k-1$ edges, then $2|E(\overline{G})| \geq f(k+1, t)$. Thus,

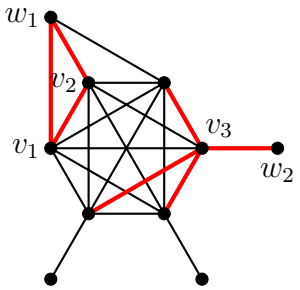
$$|E(G)| \leq \binom{k+1+t}{2} - \frac{f(k+1,t)}{2} = \binom{k+1}{2} - \frac{k+1+\mathbf{1}_{k \text{ even}}}{2}.$$

Further, if equality holds, then $G \simeq G_k \cup \overline{K_t}$, so as we do not consider graphs with isolated vertices, we must have $G \simeq G_k$. As such, G_k is the unique extremal graph for $\mathcal{E}_k(\{K_3, P_3\})$. □

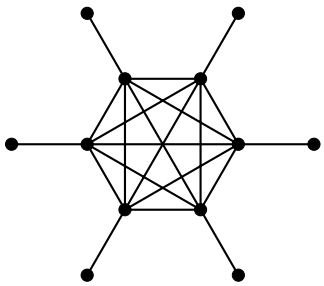
We now turn out attention to determining $\mathcal{E}_k(P_3)$. We note that H is P_3-free if and only if H is the vertex-disjoint union of triangles, stars and isolated vertices. The following graphs will be important in determining $\mathcal{E}_k(P_3)$ and classifying the extremal graphs.

Definition 4.2.15. For fixed positive integers $k, r_1, r_2, \ldots, r_s$ with $\sum_{i=1}^{s} r_i = k$, define the *pendant* graph $K_k^*(r_1, \ldots, r_s)$ as follows. Take a clique on some k-vertex set $\{v_1, \ldots, v_k\}$, the *core*, and additional vertices $\{w_1, \ldots, w_s\}$, called the *pendants*. Partition $\{v_1, \ldots, v_k\} = W_1 \cup \cdots \cup W_s$ where $|W_i| = r_i$ and connect w_i to the vertices in W_i. See Figure 4.4. Thus, the degree sequence of $K_k^*(r_1, \ldots, r_s)$ is $(\underbrace{k, \ldots, k}_{k}, r_1, \ldots, r_s)$ and $|E(K_k^*(r_1, \ldots, r_s)| = \binom{k+1}{2}$.

Lemma 4.2.16. *Let $k \geq 4$ and let $r_1, \ldots, r_s$ be positive integers with $\sum_{i=1}^{s} r_i = k-1$. We have*
$\mathrm{ex}(K_{k-1}^*(r_1, \ldots, r_s), P_3) \geq k-1$*, where equality holds if and only if either $r_i = 1$ for all i, or $3 \nmid k$ and $r_1 = k-1$. In particular, $\mathcal{E}_k(P_3) \geq \binom{k}{2}$.*



(a) $\mathrm{ex}(K_6^*(3,1,1,1), P_3) \geq 7$.



(b) $K_6^*(1,1,1,1,1,1)$.

Figure 4.4: Examples of pendant graphs.

Proof. Every vertex in the core of $G := K_{k-1}^*(r_1, \ldots, r_s)$ has degree $k-1$, so $\mathrm{ex}(G, P_3) \geq k-1$ is immediate by taking any star centered at a core vertex of G.

Now, if $r_1 = k-1$, then $G \simeq K_k$, and it is well-known that $\mathrm{ex}(K_k, P_3) = k-1$ if $3 \nmid k$. If $(r_1, \ldots, r_s) = (1, \ldots, 1)$, then let U denote the core of G. Now let $H \subseteq G$ be any P_3-free subgraph, so H is a vertex-disjoint union of triangles, stars and isolated vertices. Now, no triangle T in H can contain a pendant vertex, so each $V(T) \subseteq U$, and every star contains at most one; hence $|V(S) \cap U| \geq |V(S)| - 1$ for each star S. Hence, splitting up H into components:

$$\begin{aligned}
|E(H)| &= \sum_{\substack{T \subseteq H \\ T \text{ triangle}}} |E(T)| + \sum_{\substack{S \subseteq H \\ S \text{ star}}} |E(S)| \\
&= \sum_{\substack{T \subseteq H \\ T \text{ triangle}}} |V(T)| + \sum_{\substack{S \subset H \\ S \text{ star}}} \big(|V(S)| - 1\big) \\
&\leq \sum_{\substack{T \subseteq H \\ T \text{ triangle}}} |V(T) \cap U| + \sum_{\substack{S \subseteq H \\ S \text{ star}}} |V(S) \cap U| \leq |U| = k-1.
\end{aligned}$$

As such, $\mathrm{ex}(G, P_3) = k-1$. In particular, $\mathcal{E}_k(P_3) \geq |E(K_{k-1}^*(1, \ldots, 1)| = \binom{k}{2}$.

We now wish to show that if $G := K_{k-1}^*(r_1, \ldots, r_s)$ where $(r_1, \ldots, r_s)$ is niether $(k-1)$ nor $(1, \ldots, 1)$, then $\mathrm{ex}(G, P_3) \geq k$. Suppose that $r_1 \geq \cdots \geq r_s$, so $r_1, s \geq 2$. Let w_1, w_2 be the corresponding pendant vertices with degrees r_1, r_2, respectively. Let $v_1, v_2 \in U$ be adjacent to w_1 and let $v_3 \in U$ be adjacent to w_2 (so v_1, v_2, v_3 are distinct). Consider the graph $H \subseteq G$ which consists of the triangle w_1, v_1, v_2 and the largest star centered at v_3 which does not include v_1, v_2 (see Figure 4.4a). As $\deg(v_3) = k-1$, H is the vertex-disjoint union of a triangle and a star with $k-3$ edges. In particular, H is P_3-free, so $\mathrm{ex}(G, P_3) \geq |E(H)| = k$. □

Before determining $\mathcal{E}_k(P_3)$ exactly and classifying all extremal graphs, it is illustrative to see a small case.

Proposition 4.2.17. $\mathrm{ex}(G, P_3) = 2$ *if and only if* $G \in \{P_3, C_4\}$. *Hence,* $\mathcal{E}_3(P_3) = 4 = \binom{3}{2}+1$.

Proof. Certainly $\mathrm{ex}(P_3, P_3) = \mathrm{ex}(C_4, P_3) = 2$.

If $\mathrm{ex}(G, P_3) = 2$, then every set of 3 edges in G forms a copy of P_3. Thus, $\Delta(G) \leq 2$, G is connected and $|V(G)| \geq 4$, so G is a cycle or a path. Both P_{n-1} and C_n contain a copy of $P_1 \cup P_2$, which is P_3-free, for $n \geq 5$, so we must have $|V(G)| = 4$. As such $G \in \{P_3, C_4\}$. Thus, $\mathcal{E}_3(P_3) = 4$. □

With this out of the way, we can now completely determine $\mathcal{E}_k(P_3)$. Unfortunately, there is a fair amount of case-work involved in the proof of this theorem in order to establish the base case for an induction. For this, we turn to NAUTY to do an exhaustive search subject to the parameters which we will establish in the following proof. As is mentioned in the proof, details about this case check can be found in Appendix **??**.

Theorem 4.2.18. *For $k \geq 3$, if G is a graph with $\mathrm{ex}(G, P_3) < k$, then $|E(G)| \leq \binom{k}{2} + \mathbf{1}_{k=3}$. Furthermore, we have equality if and only if one of the following holds:*

- *$k = 3$ and $G \simeq C_4$,*
- *$k = 4$ and $G \simeq K_{2,3}$,*
- *$k \geq 4$ and $G \simeq K^*_{k-1}(1, 1, \ldots, 1)$, or*
- *$k \geq 4$, $3 \nmid k$ and $G \simeq K_k$.*

Hence, $\mathcal{E}_k(P_3) = \binom{k}{2} + \mathbf{1}_{k=3}$ for $k \geq 3$.

Proof. We first note that $\mathrm{ex}(K_{2,3}, P_3) = 3$ and $|E(K_{2,3})| = 6 = \binom{4}{2}$. Thus, along with Lemma 4.2.16 and Proposition 4.2.17, all lower bounds have been established. Additionally, Proposition 4.2.17 establishes the theorem when $k = 3$, so we will suppose $k \geq 4$ for the remainder of the proof. In fact, the small cases $4 \leq k \leq 6$ are omitted, despite constituting a cumbersome case-search, but are highly amenable to computer searches in e.g. NAUTY, once a small upper bound on $|E(G)|$ has been established. We also note that trivially, $\mathcal{E}_1(P_3) = 0 = \binom{1}{2}$ and $\mathcal{E}_2(P_3) = 1 = \binom{2}{2}$.

As such, let G be a graph with $\mathrm{ex}(G, P_3) < k$ with $|E(G)| \geq \binom{k}{2}$ and proceed by strong induction on k. Note that $\mathrm{ex}(G, P_3) \geq \Delta := \Delta(G)$, so $\Delta \leq k - 1$.

Firstly, suppose G contains a triangle $T = xyz$. If $H \subseteq G[V \setminus T] =: G'$ is P_3-free, then $H \cup T$ is also P_3-free, so $\mathrm{ex}(G', P_3) < k - 3$. Thus, by induction, $|E(G')| \leq \binom{k-3}{2} + \mathbf{1}_{k-3=3}$. Now, as $\Delta \leq k - 1$, x, y, z all have at most $k - 3$ neighbors outside T, so

$$|E(G)| \leq |E(G[V \setminus T])| + 3(k-3) + 3 \leq \binom{k-3}{2} + \mathbf{1}_{k-3=3} + 3k - 6 = \binom{k}{2} + \mathbf{1}_{k-3=3},$$

Using these facts, the cases $4 \leq k \leq 6$ can be checked exhaustively by computer search. Thus, we assume $k \geq 7$, so $|E(G)| \leq \binom{k}{2}$. If equality holds, then all of x, y, z must have

exactly $k-3$ neighbors outside of T and G' must be one of the claimed extremal graphs, so $G' \simeq K^*_{k-4}(1,\dots,1)$, or $G' \simeq K_{k-3}$ and $3 \nmid k$, or $G' \simeq K_{2,3}$ and $k-3=4$, possibly with isolated vertices.

We first consider the case where $G' \simeq K_{2,3}$, possibly with isolated vertices. In fact, we may suppose that for *every* triangle $T' \subseteq G$, we have $G[V \setminus T'] \simeq K_{2,3}$, possibly with isolated vertices, or else we may proceed as in the remaining cases. Let the vertices of the $K_{2,3}$ in G' have parts A, B where $|A|=2$ and $|B|=3$. We first note that each $v \in T$ must have all remaining $k-3=4$ edges to $A \cup B$, or else there is a $K_{1,5}$ centered at v which is disjoint from some copy of P_2 in $A \cup B$, yielding $\mathrm{ex}(G, P_3) \geq 7$; a contradiction. In particular G' has no isolated vertices. Additionally, all vertices in T must be connected to at least one vertex in A; thus, by pigeonhole, there are two vertices in T adjacent to the same vertex of A, say $y, z \sim a$. Taking $T' = yza$ shows that $G[V \setminus T'] \simeq K_{2,3}$. As such, x must be adjacent to every vertex in B and also adjacent to a. In particular, xab is a triangle for $b \in B$, so $G'' = G[V \setminus xa_1b] \simeq K_{2,3}$. However, $y \sim z$ and $\deg_{G'}(y), \deg_{G'}(z) \geq 2$, which is impossible.

Next, suppose that $G' \simeq K^*_{k-4}(1,\dots,1)$, possibly with isolated vertices, and let U denote the core of G'. If x is not adjacent to some vertex of U, then x has at least $(k-3)-(k-5)=2$ neighbors outside of $T \cup U$, denote two of these neighbors by a, b. As $|U| = k-4 \geq 3$, there must be some $u \in U$ which is not adjacent to a, b, so u is the center of a $(k-4)$-edge star in G' which does not include a, b. Thus, consider the graph $H \subseteq G$ consisting of this star centered at u along with the star $\{xy, xz, xa, xb\}$. H is P_3-free, so $\mathrm{ex}(G, P_3) \geq |E(H)| = k$; a contradiction. Hence, by symmetry, x, y, z are adjacent to all vertices in U. Thus, G is a pendant graph with core $T \cup U$. Thus, G is determined to be $K^*_{k-1}(1,\dots,1)$ by Lemma 4.2.16.

Finally, suppose $3 \nmid k$ and $G' \simeq K_{k-3}$, possibly with isolated vertices, and write $S \subseteq V \setminus T$ for the vertex set of this K_{k-3}. We notice that if x has at most one neighbor in S, then there is a star centered at x with at least $(k-1)-1 = k-2 \geq 5$ edges in G which is disjoint from S. Thus, letting H consist of a $(k-4)$-edge star in G' along with this star centered at x gives $\mathrm{ex}(G, P_3) \geq |E(H)| \geq k+1$; a contradiction. Thus, by symmetry, all of x, y, z each have at least two neighbors in S. Now, suppose that there is some $a \in V \setminus (T \cup S)$ that is adjacent to x. If $k \equiv 2 \pmod 3$, then as y has at least two neighbors in S, then we can partition $S \cup \{y\}$ into $(k-3)+1 = k-2$ vertex-disjoint triangles. Letting H consist of these triangles along with the star $\{xz, xa\}$ yields a P_3-free subgraph of G with k edges; a contradiction. Thus, suppose $k \equiv 1 \pmod 3$. Either y and z share a common neighbor in S or they each have two distinct neighbors in S. In either case, we can partition $S \cup \{y, z\}$ into $(k-3)+2 = k-1$ vertex-disjoint triangles, so letting H consist of these triangles along with the edge xa yields a P_3-free subgraph of G with k edges; another contradiction. Hence, by symmetry, x, y, z have no neighbors outside of $S \cup T$, so, in fact, $G \simeq K_k$.

After all of this, we have established the theorem if G contains a triangle, so we may suppose that G is triangle-free. As such, if $xy \in E(G)$, then $N(x) \cap N(y) = \varnothing$. Taking maximal stars with centers x and y (except for the edge xy), yields a P_3-free subgraph of G, so $k > \mathrm{ex}(G, P_3) \geq (\deg(x)-1) + (\deg(y)-1)$, so $\deg(x) + \deg(y) \leq k+1$ for every edge xy.

If there is some edge xy with $\deg(x) + \deg(y) \leq k$, then setting $G' := G \setminus \{x, y\}$ has

$|E(G')| \geq \binom{k}{2} - (k-1) = \binom{k-1}{2}$. Additionally, adding the edge xy to any P_3-free subgraph of G' shows that $\mathrm{ex}(G', P_3) \leq \mathrm{ex}(G, P_3) - 1 < k - 1$. Thus, by the induction and the fact that G' is triangle-free, we must have $k \in \{4, 5\}$ and $|E(G')| = \binom{k}{2}$. Again, we can check these cases by hand or by software.

Hence, we may suppose $\deg(x) + \deg(y) = k + 1$ for every $xy \in E(G)$. Fix x and suppose first that $d := \deg(x) \neq \frac{k+1}{2}$. Letting C denote the connected component of G containing x, we can partition $C = A \cup B$ where $A = \{u : \deg(u) = d\}$ and $B = \{u : \deg(u) = k + 1 - d\}$. As $\deg(u) + \deg(v) = k+1$ for every $uv \in E(G)$ and $d \neq \frac{k+1}{2}$, $G[C]$ is a bipartite graph with parts A, B. Now, for any $u \in A$ and $v \in B$, by considering stars centered at u and v (except for the edge uv if it exists), we find

$$k > \mathrm{ex}(G, P_3) \geq \mathrm{ex}(G[C], P_3) + \mathrm{ex}(G[V \backslash C], P_3) \geq |N(u) \backslash \{v\}| + |N(v) \backslash \{u\}| = k + 1 - 2 \cdot \mathbf{1}_{uv \in E(G)}.$$

From this, we immediately find that $G[V \setminus C]$ is empty, and as the above holds for any u, v, we know that $G[C]$ is a complete bipartite graph. Further, as C is a connected component of G and we supposed G has no isolated vertices, we have $G \simeq K_{d,k+1-d}$. Thus $|E(G)| = d(k + 1 - d) \leq \binom{k}{2}$. However, we already know that $|E(G)| \geq \binom{k}{2}$ by assumption, so $d(k + 1 - d) = \binom{k}{2}$. As $k \geq 4$, the only way for this to happen is if $k = 4$ and $d \in \{2, 3\}$. Thus, $G \simeq K_{2,3}$.

Otherwise, G is $d := (\frac{k+1}{2})$-regular. Fix $x \in V$ and set $G' := G - (N(x) \cup \{x\})$. Thus, it is clear that $\mathrm{ex}(G', P_3) + d \leq \mathrm{ex}(G, P_3) < k$, so $\mathrm{ex}(G', P_3) < k - d = \frac{k-1}{2}$. Setting $k' := \frac{k-1}{2}$, we have that $|E(G')| \leq \binom{k'}{2} + \mathbf{1}_{k'=3}$ by induction. Further, as G is triangle-free, $N(x)$ spans no edges, so

$$\binom{k'}{2} + \mathbf{1}_{k'=3} \geq |E(G')| = |E(G)| - d^2 \geq \binom{k}{2} - d^2,$$

so

$$d^2 \geq \binom{k}{2} - \binom{k'}{2} - \mathbf{1}_{k'=3} = \frac{3}{8}(k^2 - 1) - \mathbf{1}_{k'=3}$$

As k must be odd and $k \geq 4$, this is only possible if $k = 5$. Setting $k = 5$, all above inequalities become equalities, so we get $d = 3$ and $|E(G)| = \binom{5}{2}$. Thus, G is a 3-regular graph on 10 edges; an impossibility. □

The last small graph we will consider is $P_1 \cup P_2$. Determining $\mathcal{E}_k(P_1 \cup P_2)$ will also allow us to completely classify those families of graphs with $\mathcal{E}_k(\mathcal{H}) = \infty$, which we will do at the end of this section. As above, it will be important to have a complete classification of $(P_1 \cup P_2)$-free graphs.

Lemma 4.2.19. *A graph H is $(P_1 \cup P_2)$-free if and only if one of the following holds:*

- $H \simeq sK_2$ *for some s,*
- $H \simeq K_{1,s}$ *for some s,*
- $H \subseteq K_4$.

Proof. Let F be the line graph of H (whereby $V(F) := E(H)$ and $e_1 \sim_F e_2$ if and only if e_1 and e_2 share a vertex). As H is $(P_1 \cup P_2)$-free, for 3 distinct edges $e_1, e_2, e_3 \in E(H)$, then if $e_1 \nsim_F e_2$ and $e_2 \nsim_F e_3$, then it must be the case that $e_1 \nsim_F e_3$. In particular, the relation $\{(x, y) \in V(F)^2 : x = y \text{ or } x \nsim_F y\}$ is an equivalence relation on $V(F)$, so we may color $V(F) = E(H)$ so that any color class is a matching, and any two edges of a distinct color are incident.

- Suppose some color class has $s \geq 3$ edges. Since these s edges are disjoint, no other edge can be simultaneously incident to all of these, so every other color class must be empty. Thus $H \simeq sK_2$.
- Suppose some color class has 2 edges. Then all other edges must be incident to both of these, so $H \subseteq K_4$.
- Otherwise, there is 1 edge in each color, and they are all pairwise incident, so $H \simeq K_3$ or $H \simeq K_{1,s}$ for some s.

Conversely, all of these graphs are clearly $(P_1 \cup P_2)$-free. □

With this classification, determining $\mathrm{ex}(G, P_1 \cup P_2)$ for any graph G is straightforward.

Corollary 4.2.20. *For any G, $\mathrm{ex}(G, P_1 \cup P_2) = \max\{\Delta(G), M(G)\} =: t$, provided $t \geq 6$. Here $M(G)$ is the size of a maximum matching in G.*

Proof. As any star in G is $(P_1 \cup P_2)$-free, certainly $\mathrm{ex}(G, P_1 \cup P_2) \geq \Delta(G)$. Similarly, $\mathrm{ex}(G, P_1 \cup P_2) \geq M(G)$ as any matching in G is also $(P_1 \cup P_2)$-free.

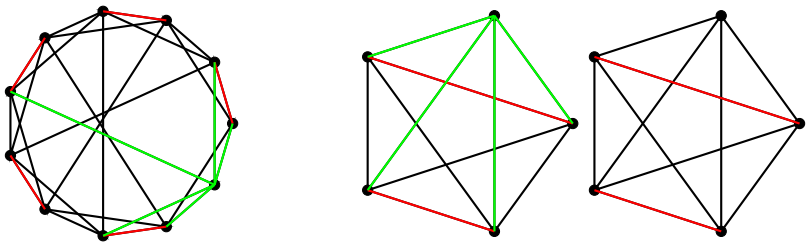
Conversely, take any subgraph $H \subseteq G$ with $t+1 > 6$ edges, so $H \not\subseteq K_4$. By the definition of t, H is neither a star nor a matching, so by Lemma 4.2.19, H must contain a copy of $P_1 \cup P_2$. Therefore, $\mathrm{ex}(G, P_1 \cup P_2) \leq t$. □

Using the above Corollary, we can provide lower bounds on $\mathcal{E}_k(P_1 \cup P_2)$.

Corollary 4.2.21. *If $k \geq 7$, then* $\mathcal{E}_k(P_1 \cup P_2) \geq \begin{cases} k^2 - \frac{3}{2}k & \text{if } k \text{ is even;} \\ k^2 - k & \text{if } k \text{ is odd.} \end{cases}$

Proof. See Figure **??** for the First suppose k is odd and consider $G := 2K_k$, so $\Delta(G) = M(G) = k - 1$. Therefore $\mathrm{ex}(G, P_1 \cup P_2) < k$ by Corollary 4.2.20 as $k \geq 7$, so $\mathcal{E}_k(P_1 \cup P_2) \geq |E(G)| = 2\binom{k}{2} = k^2 - k$.

Meanwhile, if k is even, start with the Cayley graph $H := \mathrm{Cay}\left(\mathbb{Z}_{2k-1}, \left[-\frac{k-2}{2}, \frac{k-2}{2}\right] \setminus \{0\}\right)$; that is $V(H) = \mathbb{Z}_{2k-1}$ and $xy \in E(H)$ if and only if $x - y \pmod{2k-1} \in \left[-\frac{k-2}{2}, \frac{k-2}{2}\right] \setminus \{0\}$. Now, look at all pairs of the form $\{xy : |x - y| = k/2\}$. Since $k/2$ and $2k - 1$ are coprime,



$k = 6$ $k = 5$

Figure 4.5: The largest graphs with $\mathrm{ex}(G, P_1 \cup P_2) = k$

these pairs form a Hamilton cycle in the complete graph on $\mathbb{Z}_{2k-1}$, so take any matching M among them of size $k-1$. Finally, consider the graph $G := (\mathbb{Z}_{2k-1}, E(H) \cup M)$. As M and $E(H)$ are disjoint, every vertex of G has degree $(k-1)$ with the exception of one vertex, which has degree $k-2$. Also, $M(G) = k-1$, so $\mathrm{ex}(G, P_1 \cup P_2) < k$ again by Corollary 4.2.20. Therefore,

$$\mathcal{E}_k(P_1 \cup P_2) \geq |E(G)| = \frac{1}{2} \sum_{v \in V(G)} \deg(v) = \frac{1}{2}\big((2k-2)(k-1) + (k-2)\big) = k^2 - \frac{3}{2}k. \quad \square$$

To yield upper bounds on $\mathcal{E}_k(P_1 \cup P_2)$, we prove a general bound on the number of edges of a graph based on its maximum degree and matching number. A similar theorem was proved by Abbot, Hanson and Sauer [1] in the context of the Erdős-Rado sunflower lemma, but we provide a full proof for completeness.

Theorem 4.2.22. *For a graph G, $|E(G)| \leq (\Delta(G)+1)M(G)$. Furthermore, the inequalities in Corollary 4.2.21 are in fact equalities.*

In order to prove this, we will need the following proposition, which is an immediate consequence of the Gallai-Edmonds decomposition of a graph (c.f. [40] pp. 93–95).

Proposition 4.2.23. *If G is a connected graph with the property that for every $v \in V$, $M(G-v) = M(G)$, then G has an odd number of vertices and $M(G) = \frac{|V(G)|-1}{2}$.*

Proof of Theorem 4.2.22. Let G be a graph with $M(G) \leq M$ and $\Delta(G) \leq \Delta$. Suppose G has components $S_1, \ldots, S_s, H_1, \ldots, H_t$, where S_i is a star of degree at most Δ. We will consider a series of reductions of G that maintain the matching and degree restrictions and not decrease the number of edges. We first claim that for each i and any $v \in V(H_i)$, we may suppose that $M(H_i - v) = M(H_i)$. To see this, suppose that this is not the case for some i and v. In this case, let G' be the graph formed by replacing H_i with $H_i' = H_i - v$ and adding a copy of $K_{1,\Delta}$. As $\deg(v) \leq \Delta$, we have $\Delta(G') = \Delta$ and $|E(G')| \geq |E(G)|$. Further, as every maximum matching in H_i used v, $M(H_i') = M(H_i) - 1$, so as $M(K_{1,\Delta}) = 1$, we have $M(G') = M(G) \leq M$. Thus, we may assume that $M(H_i - v) = M(H_i)$ for all i and $v \in V(H_i)$.

As such, $|V(H_i)|$ is odd and $M(H_i) = \frac{|V(H_i)|-1}{2}$ by Proposition 4.2.23. We now claim that we may suppose that $|V(H_i)| \geq \Delta + 1$ for all i. If not, form G' by replacing H_i with a copy of $\frac{|V(H_i)|-1}{2}K_{1,\Delta}$. Clearly $\Delta(G') = \Delta$, and $M(G') = M(G)$ by the previous comment. Finally,

$$|E(G')| - |E(G)| = \frac{|V(H_i)|-1}{2}\Delta - |E(H_i)| \geq \frac{|V(H_i)|-1}{2}|V(H_i)| - \binom{|V(H_i)|}{2} = 0,$$

so we may suppose this property of G. Additionally, as $|V(H_i)|$ is odd, this property tells us $|V(H_i)| \geq \Delta + 1 + \mathbf{1}_{\Delta \text{ odd}}$.

Now,

$$M \geq M(G) = s + \frac{1}{2}\sum_{i=1}^{t}\big(|V(H_i)| - 1\big),$$

so we find

$$t \leq \left\lfloor \frac{2M}{\min_i\{|V(H_i)|-1\}} \right\rfloor \leq \left\lfloor \frac{2M}{\Delta + \mathbf{1}_{\Delta \text{ odd}}} \right\rfloor.$$

Rewriting the above equation as $s + \frac{1}{2}\sum_{i=1}^{t}|V(H_i)| \leq M + t/2$, we calculate

$$\begin{aligned} |E(G)| &= \sum_{i=1}^{s}|E(S_i)| + \sum_{i=1}^{t}|E(H_i)| \\ &\leq s\Delta + \frac{\Delta}{2}\sum_{i=1}^{t}|V(H_i)| \\ &\leq \Delta\left(M + \frac{t}{2}\right) \\ &\leq \Delta\left(M + \frac{1}{2}\left\lfloor \frac{2M}{\Delta + \mathbf{1}_{\Delta \text{ odd}}} \right\rfloor\right) \\ &\leq \big(\Delta + 1\big)M. \end{aligned} \tag{4.3}$$

Now, take any $k \geq 7$ and let G be a graph with $\mathrm{ex}(G, P_1 \cup P_2) < k$, so we must have $\Delta(G), M(G) \leq k-1$. Now, when k is odd, immediately $|E(G)| \leq (\Delta(G)+1)M(G) \leq k(k-1)$; hence $\mathcal{E}_k(P_1 \cup P_2) = k^2 - k$.
When k is even, note that either $\Delta \leq k-2$, in which case immediately $|E(G)| \leq (k-1)^2 < k^2 - \frac{3}{2}k$, or else $\Delta = k-1$, so by Equation (4.3),

$$|E(G)| \leq (k-1)\left((k-1) + \frac{1}{2}\left\lfloor \frac{2(k-1)}{k} \right\rfloor\right) = (k-1)\left(k - \frac{1}{2}\right) = k^2 - \frac{3}{2}k + \frac{1}{2}.$$

Thus as k is even, we have $|E(G)| \leq k^2 - \frac{3}{2}k$, so $\mathcal{E}_k(P_1 \cup P_2) = k^2 - \frac{3}{2}k$ in this case. □

We finally conclude this section with a classification of all families that have $\mathcal{E}_k(\mathcal{H}) = \infty$.

Corollary 4.2.24. *$\mathcal{E}_k(\mathcal{H}) \leq k(k-1)$ for any $\mathcal{H}$ not containing a star or a matching. In particular, $\mathcal{E}_k(\mathcal{H}) = \infty$ if and only if $\mathcal{H}$ contains $K_{1,s}$ or sK_2 for some s.*

Proof. Suppose G is a graph with $\mathrm{ex}(G,\mathcal{H}) < k$. As $\mathcal{H}$ does not contain a star or a matching, any star or matching in G is $\mathcal{H}$-free. Thus, $\Delta(G), M(G) \leq k-1$, so $|E(G)| \leq (\Delta(G)+1)M(G) \leq k(k-1)$. □

4.2.3 Multigraphs

As mentioned earlier, if $\mathcal{H}$ is a family of simple graphs, then $\mathcal{E}_k(\mathcal{H}) \leq \mathcal{E}_k^*(\mathcal{H})$. In fact, we conjecture the following:

Conjecture 4.2.25. *If $\mathcal{H}$ consists only of simple graphs, then $\mathcal{E}_k(\mathcal{H}) = \mathcal{E}_k^*(\mathcal{H})$.*

This statement appears very difficult to prove in general. Indeed, in [6] and [13], a similar conjecture has been put forth specifically for $\mathcal{C}_o$, the family of odd cycles, i.e. when considering max cuts (or "judicious partitions"), but in a slightly stronger setting.

However, we can present the proof of a simple subcase.

Proposition 4.2.26. *Let $\mathcal{H}$ be a family of simple graphs and G be a multigraph. If each edge of G has the same multiplicity, then there exists a simple graph G' with $|E(G')| = |E(G)|$ and $\mathrm{ex}(G',\mathcal{H}) \leq \mathrm{ex}(G,\mathcal{H})$.*

Proof. Let G be a multigraph where each edge has multiplicity r. Decompose G into simple graphs $G_1,\ldots,G_r$ and let G' be the disjoint union of $G_1,\ldots,G_r$, so certainly we have $|E(G')| = |E(G)|$. Now, let $F \subseteq G'$ be an $\mathcal{H}$-free subgraph on $\mathrm{ex}(G',\mathcal{H})$ edges and set $F_i = F \cap G_i$. Without loss of generality, suppose $|E(F_1)| \geq |E(F_i)|$ for all i and form $F' \subseteq G$ by replacing each edge of F_1 by r copies. As $\mathcal{H}$ consisted only of simple graphs, F' is also $\mathcal{H}$-free, so

$$\mathrm{ex}(G,\mathcal{H}) \geq |E(F')| = r|E(F_1)| \geq r \cdot \frac{\mathrm{ex}(G',\mathcal{H})}{r} = \mathrm{ex}(G',\mathcal{H}). \qquad \square$$

Unfortunately, when G is a multigraph where different edges have different multiplicities, it is unclear whether or not one can construct a simple graph G' with $|E(G')| = |E(G)|$ and $\mathrm{ex}(G',\mathcal{H}) \leq \mathrm{ex}(G,\mathcal{H})$.

Notice (see Theorem 4.3.5) that if $\mathcal{H}$ does not contain a bipartite graph, then $\mathcal{E}_k^*(\mathcal{H}) = \big(1+o(1)\big)\mathcal{E}_k(\mathcal{H})$. We can also provide the following bound which, unfortunately, is not very strong.

Proposition 4.2.27. *If $\mathcal{H}$ is a family of simple graphs, then $\mathcal{E}_k^*(\mathcal{H}) \leq \mathcal{E}_{k\log k}(\mathcal{H})$.*

Proof. Both are infinite if $\mathcal{H}$ contains a star or a matching, so we shall suppose that is not the case.

Let G be a multigraph with $\mathrm{ex}(G,\mathcal{H}) < k$. As above, decompose G into simple graphs $G_1,\ldots,G_r$ where $G_1 \supseteq \cdots \supseteq G_r$, and let G' be the disjoint union of these graphs, so

certainly $|E(G')| = |E(G)|$. We now argue that $\text{ex}(G', \mathcal{H}) < k \log k$, which will establish the claim.

To do this, we first note that as $\text{ex}(G, \mathcal{H}) < k$, we must have $r \leq k-1$ as $\mathcal{H}$ does not contain K_2. Further, consider any $\mathcal{H}$-free subgraph $F \subseteq G_i$. As $G_1 \supseteq \cdots \supseteq G_r$, and $\mathcal{H}$ is a family of simple graphs, we can form an $\mathcal{H}$-free subgraph $F' \subseteq G$ by replacing every edge of F by i copies. Thus, it must be the case that $\text{ex}(G_i, \mathcal{H}) < \frac{k}{i}$. As such,

$$\text{ex}(G', \mathcal{H}) \leq \sum_{i=1}^{r} \text{ex}(G_i, \mathcal{H}) < \sum_{i=1}^{r} \frac{k}{i} \leq k \log(r+1) \leq k \log k. \qquad \square$$

Interestingly, Conjecture 4.2.25 fails if we consider non-uniform hypergraphs, and in Section 4.3.1, we give an example of such a hypergraph.

4.3 Hypergraphs

We now explore the extremal function $\mathcal{E}_k(\mathcal{H})$ when $\mathcal{H}$ is a family of hypergraphs.

In light of the result on 2-uniform graphs, we begin by asking when $\mathcal{E}_k(\mathcal{H}) = \infty$ for a family of hypergraphs $\mathcal{H}$ of higher uniformity. In fact, this is answered by the classical sunflower lemma due to Erdős and Rado [18].

Definition 4.3.1. Let H be any r-uniform (multi)hypergraph. H is said to be a *sunflower* if, for some $S \subseteq V(H)$ called the *core* of H, *every* pair of distinct edges $e_1, e_2 \in E(H)$ has $e_1 \cap e_2 = S$. Note that $K_{1,s}$ and sK_2 fully describe all simple 2-uniform sunflowers (where $|S| = 1, 0$ respectively).

Crucially, whenever H is a sunflower, *every* $F \subseteq H$ is also a sunflower.

Proposition 4.3.2. *$\mathcal{E}_k^*(\mathcal{H}) = \infty$ for k sufficiently large if and only if $\mathcal{H}$ contains a sunflower.*

Proof. If $\mathcal{H}$ contains a sunflower H with $|E(H)| = k$ and core S, then any sunflower G with s edges and core of size $|S|$ has $\text{ex}(G, H) = k - 1$. Hence, $\mathcal{E}_k^*(\mathcal{H}) \geq s$ for every s, so $\mathcal{E}_k^*(\mathcal{H}) = \infty$.

Conversely, take any family of hypergraphs $\mathcal{H}$ without a sunflower and fix k. By the Erdős-Rado sunflower lemma [18], any r-graph G with $|E(G)| > r!(k-1)^{r+1}$ [1] contains a sunflower F with at least k edges. Thus F contains no hypergraph in $\mathcal{H}$, showing $\text{ex}(G, \mathcal{H}) \geq k$. The contrapositive gives us $\mathcal{E}_k^*(\mathcal{H}) \leq r!(k-1)^{r+1}$. $\square$

We will also show later that, for *most* uniform hypergraphs, cliques are asymptotically best at forcing them. The only possible exceptions are when the hypergraphs are "degenerate":

[1] When G is simple, this can be lowered to $r!(k-1)^r$.

Definition 4.3.3. For an arbitrary r-uniform hypergraph family $\mathcal{H}$ we denote by

$$\left(\pi_n(\mathcal{H}) := \frac{\mathrm{ex}(K_n^{(r)}, \mathcal{H})}{\binom{n}{r}}\right)_{n\geq 1}$$

the sequence of Turán densities and denote the limiting density $\pi(\mathcal{H}) := \lim_{n\to\infty} \pi_n(\mathcal{H})$.

$\mathcal{H}$ is said to be *degenerate* if $\pi(\mathcal{H}) = 0$.

Note that $(\pi_n(\mathcal{H}))_{n\geq 1}$ is a decreasing sequence of densities for any $\mathcal{H}$ by averaging over subgraphs, so the limit always exists. Furthermore, there is a standard classification:

Proposition 4.3.4. *An r-uniform graph H is degenerate if and only if it is r-partite. That is to say, we may r-color $V(H)$ so that each $e \in E(H)$ has 1 vertex of each color, or equivalently, $H \subseteq K^{(r)}_{\underbrace{t,t,\ldots,t}_{r}}$ for some t.*

Indeed, for H nondegenerate, $\pi(H) \geq r!/r^r$ as the balanced r-partite hypergraph $K^{(r)}_{n/r,\ldots,n/r} \not\supseteq H$, otherwise $\mathrm{ex}(K_n^{(r)}, H) = o(n^r)$ is true by an induction on r, as was observed by Erdős [17]. In fact, these easily generalize to families of r-uniform graphs; namely $\pi(\mathcal{H}) = 0$ if and only if $\mathcal{H}$ contains a degenerate graph. See [32] for a survey on the hypergraph Turán problem.

Theorem 4.3.5. *If $\mathcal{H}$ is a family of simple r-uniform hypergraphs not containing a degenerate graph, then*

$$\mathcal{E}_k(\mathcal{H}), \mathcal{E}_k^*(\mathcal{H}) = \left(\frac{1}{\pi(\mathcal{H})} - o(1)\right) k.$$

Proof. With the exception of applying contractions, we proceed in a fashion similar to the proof of Theorem 4.2.6.

Lower bound. For a positive integer k, let n be the largest integer for which $k > \pi_n(\mathcal{H})\binom{n}{r}$. As $\pi_n(\mathcal{H}) = \pi(\mathcal{H}) + o(1)$ and $\binom{n+1}{r} - \binom{n}{r} = O(n^{r-1})$, we observe that $\pi_{n+1}(\mathcal{H})\binom{n+1}{r} - \pi_n(\mathcal{H})\binom{n}{r} = o(n^r)$; thus $k \leq \pi_n(\mathcal{H})\binom{n}{r} + o(k)$. Then, as $\mathrm{ex}(K_n^{(r)}, \mathcal{H}) = \pi_n(\mathcal{H})\binom{n}{r} < k$,

$$\mathcal{E}_k(\mathcal{H}) \geq |E(K_n^{(r)})| = \binom{n}{r} \geq \frac{k - o(k)}{\pi_n(\mathcal{H})\binom{n}{r}}\binom{n}{r} = \left(\frac{1}{\pi(\mathcal{H})} - o(1)\right)k.$$

Upper bound. Let G be an r-uniform (multi)graph on n vertices with $\mathrm{ex}(G, \mathcal{H}) < k$, and let $F \subseteq K_n^{(r)}$ be an $\mathcal{H}$-free subgraph with $|E(F)| = \mathrm{ex}(K_n^{(r)}, \mathcal{H}) = \pi_n(\mathcal{H})\binom{n}{r}$. Let F' be a copy of F chosen uniformly at random from $K_n^{(r)}$ and set $F^* = \{e \in E(G) : e \in E(F')\}$, where multiedges are preserved. Certainly as F is $\mathcal{H}$-free and $\mathcal{H}$ consists only of simple graphs, F^* is also $\mathcal{H}$-free. Therefore, as $\pi_n(\mathcal{H}) \geq \pi(\mathcal{H})$,

$$k > \mathbf{E}|E(F^*)| = \pi_n(\mathcal{H})|E(G)| \geq \pi(\mathcal{H})|E(G)|,$$

so $\mathcal{E}_k^*(\mathcal{H}) < \frac{k}{\pi(\mathcal{H})}$. □

As degenerate 2-uniform graphs are exactly bipartite graphs, Theorem 4.3.5 immediately implies Theorem 4.1.1 by noting that $\pi(\mathcal{H}) = 1 - \frac{1}{\rho(\mathcal{H})-1}$ by the Erdős-Stone Theorem [20].

Unfortunately, when it comes to hypergraphs, we cannot attain a tighter result when $\mathcal{H} = \{K_t^{(r)}\}$ as we could in Theorem 4.2.6. The main difficulty here is that when $r \geq 3$, it may not be possible to apply compressions to end up with a clique at the end. However, despite this difficulty, it should still be the case that cliques are extremal for $\mathcal{E}_k(K_t^{(r)})$.

Conjecture 4.3.6. *If $k = \mathrm{ex}(K_n^{(r)}, K_t^{(r)})+1$, then $\mathcal{E}_k(K_t^{(r)}) = \mathcal{E}_k^*(K_t^{(r)}) = \binom{n}{r}$ and the unique extremal graph is $K_n^{(r)}$.*

To end this section, we present a general upper bound on $\mathcal{E}_k(H)$, which directly follows from the work of Friedgut and Kahn [25] who extended a result of Alon [4].

For two hypergraphs H and G, let $N(G, H)$ denote the number of copies of H contained in G, and let $N(m, H)$ denote the maximum value of $N(G, H)$ taken over all hypergraphs G, with $|E(G)| = m$. Also, for a hypergraph H, we say that $\phi : E(H) \to [0, 1]$ is a *fractional cover of H* if $\sum_{e \ni v} \phi(e) \geq 1$ for every $v \in V(H)$. The *fractional cover number of H*, denoted $\rho^*(H)$ is the minimum value of $\sum_{e \in E(H)} \phi(e)$ where ϕ is a fractional cover of H.

Theorem 4.3.7 (Friedgut and Kahn [25])**.** *For any hypergraph H, $N(m, H) = \Theta(m^{\rho^*(H)})$.*

Proposition 4.3.8. *If $\rho^* = \rho^*(H)$ and $s = |E(H)|$, then there is a constant $c = c(H)$ such that*

$$\mathcal{E}_k(H) \leq ck^{(s-1)/(s-\rho^*)}.$$

Proof. Let G be a graph with $\mathrm{ex}(G, H) < k$ and $|E(G)| = m$. Thus, by Theorem 4.3.7, there is a constant $C = C(H)$ such that $N(G, H) \leq N(m, H) \leq Cm^{\rho^*}$.

We proceed by a standard averaging argument. Let $S \subseteq E(G)$ be a set of edges where each $e \in E(G)$ is included in S independently with probability p. Then let $S' \subseteq S$ be attained by removing one edge per copy of H contained in S. Thus S' is H-free, so

$$k > \mathbf{E}|S'| \geq pm - p^s N(G, H) \geq pm - Cp^s m^{\rho^*} = pm\big(1 - Cp^{s-1}m^{\rho^*-1}\big).$$

Selecting $p^{s-1}m^{\rho^*-1} = 1/(sC)$ yields

$$k > \left(1 - \frac{1}{s}\right)\left(\frac{m^{s-\rho^*}}{sC}\right)^{1/(s-1)}.$$

As such, there is some $c = c(H)$ with

$$m < ck^{(s-1)/(s-\rho^*)}.$$

□

4.3.1 Non-uniform Hypergraphs

Recall $\text{ex}(G, \mathcal{H}) = |E(G)|$ unless G contains a copy of some $H \in \mathcal{H}$, so it makes sense to even ask about $\mathcal{E}_k^*(\mathcal{H})$ where $\mathcal{H}$ is a family of non-uniform hypergraphs.

Throughout this section, for a graph G, we will use $E_i(G) := \{e \in E(G) : |e| = i\}$.

Proposition 4.3.9. *If H is a non-uniform hypergraph, then $\mathcal{E}_k^*(H) \leq 2(k-1)$. Additionally, if $\mathcal{H}$ is a finite family of non-uniform hypergraphs, then $\mathcal{E}_k^*(\mathcal{H})$ is always finite.*

Proof. As H is non-uniform, there is some $r \neq s$ with $E_r(H), E_s(H) \neq \varnothing$. Now, let G be any hypergraph with $\text{ex}(G, H) < k$. As any $F \subseteq G$ with $E_r(F) = \varnothing$ or $E_s(F) = \varnothing$ is trivially H-free, we know that $|E_r(G)| < k$ and $|E(G) \setminus E_r(G)| < k$, therefore, $|E(G)| \leq 2(k-1)$.

Now take a finite family of non-uniform graphs $\mathcal{H}$. Let $U = \{i \in \mathbb{Z} : \exists H \in \mathcal{H}, E_i(H) \neq \varnothing\}$ be the set of all edge uniformities appearing in $\mathcal{H}$. Let G be a hypergraph with $\text{ex}(G, \mathcal{H}) < k$; certainly we may suppose that the edges in G are only of the sizes in U. Thus, by the same argument as above, $|E_i(G)| < k$ for all $i \in U$ as each $H \in \mathcal{H}$ is non-uniform, so $|E(G)| \leq |U|(k-1)$, which is finite as $\mathcal{H}$ consisted only of finitely many graphs. □

We quickly remark that $\mathcal{E}_k^*(\mathcal{H})$ is not necessarily finite when $\mathcal{H}$ is not of finite size. Namely, for positive integers r, t, let $H_{r,t}$ be the non-uniform hypergraph consisting of two disjoint edges e, s with $|e| = r$, $|s| = t$. Then $\mathcal{H} = \{H_{r,t} : 1 \leq r < t\}$ has $\mathcal{E}_k^*(\mathcal{H}) = \infty$ when $k \geq 2$, as is realized by taking a host graph with disjoint edges $e_1, \ldots, e_s$ where $|e_i| = i$.

We now turn our attention to a non-uniform hypergraph which yields a surprising answer to $\mathcal{E}_k^*(H)$; namely $\mathcal{E}_k^*(H) \sim \alpha k$ where α is an irrational number. The *r-necklace*, denoted $\mathcal{O}_r$, is the hypergraph with vertex set $\{x_1, \ldots, x_r\}$ and edge set $\big\{\{x_1, \ldots, x_r\}, \{x_1\}, \ldots, \{x_r\}\big\}$. That is, $\mathcal{O}_r$ is the hypergraph consisting of a single r-edge with a loop at each vertex.

Theorem 4.3.10. *For $r \geq 2$, $\mathcal{E}_k^*(\mathcal{O}_r) = \big(\frac{1}{\alpha_r} - o(1)\big)k$ where α_r is the unique positive solution to $X^r + X = 1$.*

Proof. Upper bound. Let G be any hypergraph with $\text{ex}(G, \mathcal{O}_r) < k$; certainly we may assume G contains only contains edges of uniformities 1 and r. Now, let $V' \subseteq V(G)$ be formed by including each vertex in V' with probability α_r, and form $G' \subseteq G$ by taking any loops on a vertex of V' along with any r-uniform edge which is not completely contained in V'. By construction, G' is $\mathcal{O}_r$-free, so

$$k > \mathbf{E}|E(G')| = \alpha_r |E_1(G)| + (1 - \alpha_r^r)|E_r(G)| = \alpha_r\big(|E_1(G)| + |E_r(G)|\big) = \alpha_r |E(G)|.$$

Therefore, $|E(G)| < k/\alpha_r$, so the same is true of $\mathcal{E}_k^*(\mathcal{O}_r)$.

Lower bound. We will show $\mathcal{E}_k^*(\mathcal{O}_r) \geq k/\alpha_r - O(k^{\frac{r}{r+1}})$.

Fix a large k, and construct the multigraph G on $n = \Theta(k^{\frac{1}{r+1}})$ vertices with:

- $t := \left\lfloor \dfrac{k}{\binom{n}{r}} \cdot \dfrac{1}{\alpha_r + r\alpha_r^r} \right\rfloor$ parallel hyperedges spanning every r-set of vertices, and
- $s := \left\lfloor \dfrac{k}{n} \cdot \dfrac{r\alpha_r^{r-1} - \frac{r^2}{n-r}}{\alpha_r + r\alpha_r^r} \right\rfloor$ loops at each vertex.[2]

This way, G has $|E(G)| = k\left(\frac{1}{\alpha_r + r\alpha_r^r} + \frac{r\alpha_r^{r-1} - \frac{r^2}{n-r}}{\alpha_r + r\alpha_r^r}\right) - O(k^{\frac{r}{r+1}}) = k/\alpha_r - O(k^{\frac{r}{r+1}})$.

Now, take any $\mathcal{O}_r$-free subgraph $H \subseteq G$. We will show $|E(H)| < k$.

Let $L \subseteq V(H)$ be the vertices of H with at least one loop. Write $\beta n := |L|$, then certainly H has at most βns loops in total. If $\beta n < r$, then we calculate

$$|E(H)| \le \beta ns + t\binom{n}{r} \le r\left(\frac{k}{n} \cdot \frac{r\alpha_r^{r-1} - \frac{r^2}{n-r}}{\alpha_r + r\alpha_r^r}\right) + \frac{k}{\alpha_r + r\alpha_r^r} = \frac{k}{\alpha_r + r\alpha_r^r} + \Theta(k^{\frac{r}{r+1}}) < k$$

for k sufficiently large as $\alpha_r + r\alpha_r^r = 1 + (r-1)\alpha_r^r > 1$, so suppose $\beta n \ge r$. In this case, we note the inequality

$$\begin{aligned}
\frac{\binom{\beta n}{r}}{\binom{n}{r}} &= \frac{\beta n}{n}\frac{\beta n - 1}{n-1}\cdots\frac{\beta n - r + 1}{n - r + 1} \\
&= \beta\left(\beta - (1-\beta)\frac{1}{n-1}\right)\left(\beta - (1-\beta)\frac{2}{n-2}\right)\cdots\left(\beta - (1-\beta)\frac{r-1}{n-r+1}\right) \\
&\ge \beta^r - \beta^{r-1}(1-\beta)\sum_{i=1}^{r-1}\frac{i}{n-i} \\
&\ge \beta^r - \beta^{r-1}(1-\beta)\frac{r^2}{n-r}.
\end{aligned}$$

We also note that $\beta \ne \alpha_r$ as β is rational and α_r is irrational; therefore, by the mean value theorem, there is some θ strictly between α_r and β such that $(\alpha_r - \beta)r\theta^{r-1} = \alpha_r^r - \beta^r$. Thus,

$$\beta r\alpha_r^{r-1} + \alpha_r + \alpha_r^r - \beta^r = \alpha_r + r(\alpha_r - \beta)(\theta^{r-1} - \alpha_r^{r-1}) + r\alpha_r^r < \alpha_r + r\alpha_r^r,$$

as either $\alpha_r < \theta < \beta$ or $\beta < \theta < \alpha_r$. Now, as H is $\mathcal{O}_r$-free, there are no r-edges spanned by L, so, noting that $\beta^{r-1}(1-\beta) \le \beta$,

$$\begin{aligned}
\frac{|E(H)|}{k} &\le \frac{\beta ns + t\left(\binom{n}{r} - \binom{\beta n}{r}\right)}{k} \\
&\le \frac{\beta\left(r\alpha_r^{r-1} - \frac{r^2}{n-r}\right) + \left(1 - \beta^r + \beta^{r-1}(1-\beta)\frac{r^2}{n-r}\right)}{\alpha_r + r\alpha_r^r}
\end{aligned}$$

[2]The constants $1/(\alpha_r + r\alpha_r^r)$ and $r\alpha_r^{r-1}/(\alpha_r + r\alpha_r^r)$ may be found by solving the natural linear program, but this is not necessary for the proof.

$$\leq \frac{\beta r\alpha_r^{r-1} + (1-\beta^r)}{\alpha_r + r\alpha_r^r}$$
$$= \frac{\beta r\alpha_r^{r-1} + \alpha_r + \alpha_r^r - \beta^r}{\alpha_r + r\alpha_r^r}$$
$$< \frac{\alpha_r + r\alpha_r^r}{\alpha_r + r\alpha_r^r} = 1. \qquad \square$$

In the case of $r = 2$, where $\mathcal{O}_2$ is a 2-uniform edge with a loop at both ends, we attain an interesting corollary.

Corollary 4.3.11. $\mathcal{E}_k^*(\mathcal{O}_2) = \big(\phi - o(1)\big)k$ *where* $\phi = 1.618\ldots$ *is the golden ratio.*

We now show that for a simple, non-uniform hypergraph H, it can be the case that $\mathcal{E}_k(H)$ and $\mathcal{E}_k^*(H)$ differ. This is perhaps surprising as we believe it should be the case that $\mathcal{E}_k(H) = \mathcal{E}_k^*(H)$ if H is a simple r-uniform graph as mentioned in Conjecture 4.2.25.

Theorem 4.3.12. *If G is a graph with 1-uniform edges and 2-uniform edges, where each vertex has at most one loop (but any 2-uniform edges can have higher multiplicity), then* $\mathrm{ex}(G, \mathcal{O}_2) \geq \frac{2}{3}|E(G)|$.

Proof. Let G be a graph with only 2-uniform edges and loops where each vertex has at most one loop. As before, let $E_i(G)$ denote the set of i-uniform edges, so $E(G) = E_1(G) \cup E_2(G)$. We begin by claiming that we may suppose that every vertex of G has a loop. If some $v \in V(G)$ does not have a loop, then either v is isolated, in which case we may simply delete v, or v is incident to some $e \in E_2(G)$. Let G' be formed by deleting e and adding a loop around v. Certainly $\mathrm{ex}(G', \mathcal{O}_2) \leq \mathrm{ex}(G, \mathcal{O}_2)$ as the edge $e \in E_2(G)$ cannot be used in any copy of $\mathcal{O}_2$. After this reduction, we know that $|E(G)| = |E_1(G)| + |E_2(G)| = |V(G)| + |E_2(G)|$. Additionally, we may suppose that every vertex is incident to some $e \in E_2(G)$. To see this, suppose $v \in V(G)$ is not incident to any edge in $E_2(G)$; pick any $e \in E_2(G)$ and form G' by removing the loop from v and adding an additional copy of the edge e. As the loop around v cannot be used in any copy of $\mathcal{O}_2$ in G, we see that $\mathrm{ex}(G', \mathcal{O}_2) \leq \mathrm{ex}(G, \mathcal{O}_2)$.

We now prove the statement by induction on $|V(G)|$.

For the base case, suppose that $E_2(G)$ is bipartite with partite sets A, B where $|A| \geq |B|$. In this case, if we take every edge in $E_2(G)$ and every loop around a vertex in A, we end up with an $\mathcal{O}_2$-free graph as no two loops are joined by an edge. Now, as G has no vertices not incident to a 2-edge, we must have $|E_2(G)| \geq |B|$, so as $|A| \geq |B|$, we have

$$\mathrm{ex}(G, \mathcal{O}_2) \geq |E_2(G)| + |A| \geq \frac{2}{3}|E(G)|.$$

Now suppose that $E_2(G)$ is not bipartite and let $C \subseteq G$ be an induced copy of C_{2t+1} for some t, possibly with some multiedges. Set $G' := G \setminus C$.

Now, for a fixed set of vertices $S \subseteq V(C)$, we may form $H_S \subseteq G$ by collecting together the following edges:

- All 2-edges in the cycle C itself (there are at least $2t+1$ of these),
- The 2-edges from $C\backslash S$ to $V\backslash C$,
- All loops in S,
- $E(H')$ for some extremal $\mathcal{O}_2$-free $H' \subseteq G'$.

Provided S contains no two adjacent vertices in C, H_S is $\mathcal{O}_2$-free.

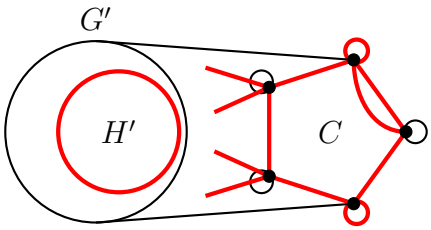


Figure 4.6: H_S edges in red, $G\backslash H_S$ edges in black (here $|S| = 2$).

So, suppose we choose $S \subseteq V(C)$ by picking an independent set of size $\lceil \frac{2t+1}{3} \rceil$ uniformly at random with probability $\frac{2t+1}{3} - \lfloor \frac{2t+1}{3} \rfloor$, otherwise an independent set of size $\lfloor \frac{2t+1}{3} \rfloor$ uniformly at random. Since $\lceil \frac{2t+1}{3} \rceil \leq t = \alpha(C_{2t+1})$, this is a nontrivial probability space. Furthermore, the event $\{v \in S\}$ occurs with probability $\frac{1}{3}$ for each $v \in C$.

Recalling that by induction, $|E(H')| = \mathrm{ex}(G', \mathcal{O}_2) \geq \frac{2}{3}|E(G')|$, we have in total

$$\begin{aligned}\mathbf{E}[|E(H_S)|] &= |E(C)| + \frac{2}{3}|E(G[C, V\backslash C])| + \frac{1}{3}(2t+1) + \frac{2}{3}|E(G')| \\ &\geq \frac{2}{3}|E(C)| + \frac{2}{3}|E(G[C, V\backslash C])| + \frac{2}{3}(2t+1) + \frac{2}{3}|E(G')| = \frac{2}{3}|E(G)|.\end{aligned}$$

So some such S yields an $\mathcal{O}_2$-free H_S with at least this many edges, as desired. □

Thus, we have the following corollary which shows that Conjecture 4.2.25 can fail for non-uniform graphs.

Corollary 4.3.13. *$\mathcal{E}_k(\mathcal{O}_2) < \frac{3}{2}k$ whereas $\mathcal{E}^*_k(\mathcal{O}_2) = \big(\phi - o(1)\big)k \approx \big(1.618 - o(1)\big)k$.*

4.3.2 1-Uniform Graphs

A 1-uniform graph on n vertices is equivalent to its degree sequence $(d_1, \ldots, d_n)$ where d_i is the number of loops at vertex i. For 1-uniform graphs $H = (d_1, \ldots, d_n)$ and $G = (x_1, \ldots, x_t)$, $H \subseteq G$ if and only if there is an injection $f : [n] \to [t]$ such that for every $i \in [n]$, $d_i \leq x_{f(i)}$.

We quickly note that a 1-uniform graph H is a sunflower if and only if it is of the form $H = (1, 1, \ldots, 1)$ or $H = (r)$ for some r.

Although the Turán problem for 1-uniform graphs is quite uninteresting as every simple 1-uniform graph is a sunflower, determining $\mathcal{E}_k^*(H)$ requires some more thought. One reason for caring about 1-uniform graphs in this context is that it also settles the question for multi-stars. For positive integers $d_1, \dots, d_t$, the multi-star $S_{d_1,\dots,d_t}$ is a star on $t+1$ vertices whose edges have multiplicities $d_1, \dots, d_t$.

Observation 4.3.14. *For positive integers $d_1, \dots, d_t$, if $H = (d_1, \dots, d_t)$, then $\mathcal{E}_k^*(S_{d_1,\dots,d_t}) = \mathcal{E}_k^*(H)$.*

Theorem 4.3.15. *For every 1-uniform hypergraph $H = (d_1, d_2, \dots, d_t)$ with $d_1 \geq \cdots \geq d_t \geq 1$ where $d_1, t \geq 2$, there exists a constant c_H such that $\mathcal{E}_k^*(H) = \big(c_H + o(1)\big)k^2$. Additionally, c_H can be determined in polynomial time and satisfies $\frac{1}{4(t-1)(d_1-1)} \leq c_H \leq \frac{1}{(t-1)(d_1-1)}$.*

Proof. Let $H = (d_1, \dots, d_t)$ where $d_1 \geq \cdots \geq d_t \geq 1$ and $d_1, t \geq 2$.

We note that $F \subseteq G$ with $F = (f_1, \dots, f_n)$ can be assumed to have $f_1 \geq \cdots \geq f_n$. Thus, it is clear that F is H-free if and only if there is some $t' \in [t]$ such that $f_{t'} < d_{t'}$. Thus, for $t' \in [t]$, let $G_{t'} = (x'_1, \dots, x'_n)$ where $x'_i = x_i$ for $i < t'$ and $x'_i = \min\{x_i, d_{t'} - 1\}$ for all $i \geq t'$. By the earlier note, $G_{t'}$ is H-free for every $t' \in [t]$, and further, any $F \subseteq G$ that is H-free must be contained in some $G_{t'}$. Thus,

$$\mathrm{ex}(G, H) = \max_{t' \in [t]} |E(G_{t'})|.$$

As $H \neq (1, 1, \dots, 1)$, we know that if $\mathrm{ex}(G, H) < k$, then $n \leq k-1$. Thus, we may formulate the following non-linear integer program for $\mathcal{E}_k^*(H)$:

$$\begin{array}{rcll} \mathcal{E}_k^*(H) & = & \max & \sum_{i=1}^{k-1} x_i & \\ & & \text{s.t.} & \sum_{i=1}^{t'-1} x_i + \sum_{i=t'}^{k-1} \min\{x_i, d_{t'} - 1\} \leq k-1 & \text{for all } t' \in [t] \\ & & & x_i \in \mathbb{Z}_{\geq 0} & \text{for all } i \in [k-1]. \end{array}$$

Fix a feasible G. Note that, since $t \geq 2$, taking $t' = 2$ shows $x_1 \leq x_1 + \sum_{i=2}^{k-1} \min\{x_i, d_2 - 1\} \leq k-1$.

Now, define $j := \max\{i : x_i \geq d_1\}$. Then $\sum_{i>j} x_i \leq (k-1)(d_1 - 1) \leq d_1 k$. Furthermore, if the largest j vertices $(x_1, \dots, x_j)$ differ in degree by ≥ 2, then certainly $x_i \geq x_{i+1} + 1 \geq \cdots \geq x_{\ell-1} + 1 \geq x_\ell + 2$ for some $i < \ell \leq j$.

Then forming G' by replacing x_i, x_ℓ with $x_i - 1, x_\ell + 1$ respectively (noting the degree sequence is still decreasing) is still feasible, for otherwise the first condition is violated for some $t' \in [t]$. This would mean

$$\min\{x_i - 1, d_{t'} - 1\} + \min\{x_\ell + 1, d_{t'} - 1\} > \min\{x_i, d_{t'} - 1\} + \min\{x_\ell, d_{t'} - 1\},$$

as only x_i and x_ℓ changed in value when forming G'. Thus $\min\{x_\ell + 1, d_{t'} - 1\} = x_\ell + 1$ so $x_\ell \leq d_{t'} - 2 < d_1$; a contradiction.

Thus, we may suppose $G = (x_1, \ldots, x_n)$ where $|x_i - x_\ell| \leq 1$ for all $i, \ell \leq j$. From this, define $G^{(1)} := (\underbrace{x_j, x_j, \ldots, x_j}_{j}, 0, \ldots, 0)$, which is also feasible and has

$$|E(G)| - |E(G^{(1)})| = \sum_{i=1}^{j}(x_i - x_j) + \sum_{i=j+1}^{k-1} x_i \leq j + d_1 k = O(k).$$

As such, we have $f^{(1)}(H) \leq \mathcal{E}_k^*(H) \leq f^{(1)}(H) + O(k)$ where

$$\begin{array}{rcll} f_k^{(1)}(H) & = & \max & jx \\ & & \text{s.t.} & (t'-1)x + \sum_{i=t'}^{j} \min\{x, d_{t'} - 1\} \leq k-1 \quad \text{for all } t' \in [t] \\ & & & x, j \in \mathbb{Z}_{\geq 0},\ x \geq d_1, \end{array}$$

where the lower bound follows from the fact that for a feasible pair (x, j), the 1-graph $(\underbrace{x, x, \ldots, x}_{j}, 0, \ldots, 0)$ satisfies the original program.

To simplify further, note that $x > d_{t'} - 1$ for any $t' \in [t]$ as $x \geq d_1$, so we know that $\min\{x, d_{t'} - 1\} = d_{t'} - 1$. Further, if a feasible (x, j) has $j < t$, then the objective is $xj < (k-1)t = O(k)$. Whether or not the optimum is among such (x, j), this shows we decrease the objective by at most $O(k)$ upon imposing the restriction $j \geq t$. Thus,

$$\begin{array}{rcll} f_k^{(2)}(H) & = & \max & jx \\ & & \text{s.t.} & (t'-1)x + (j - t' + 1)(d_{t'} - 1) \leq k-1 \quad \text{for all } t' \in [t] \\ & & & x \geq d_1 \\ & & & j \geq t \\ & & & x, j \in \mathbb{Z}, \end{array}$$

has $f_k^{(2)}(H) \leq f_k^{(1)}(H) \leq f_k^{(2)}(H) + O(k)$.

Next, replace j with $j - t$, and x with $x - d_1$, noting that the objective function decreases by $xj - (x - d_1)(j - t) \leq xt + jd_1 \leq O(k)$. Thus

$$\begin{array}{rcll} f_k^{(3)}(H) & = & \max & jx \\ & & \text{s.t.} & (t'-1)x + (d_{t'} - 1)j \leq k-1 \quad \text{for all } t' \in [t] \\ & & & x, j \in \mathbb{Z}_{\geq 0} \end{array}$$

satisfies $f_k^{(3)}(H) \leq f_k^{(2)}(H) \leq f_k^{(3)}(H) + O(k)$.

We now relax the integrality of x, j to attain

$$\begin{array}{rcll} f_k^{(4)}(H) & = & \max & jx \\ & & \text{s.t.} & (t'-1)x + (d_{t'} - 1)j \leq k-1 \quad \text{for all } t' \in [t] \\ & & & x, j \geq 0 \end{array}$$

and note that as $xj - \lfloor x \rfloor \lfloor j \rfloor \leq x + j = O(k)$, we have $f_k^{(4)}(H) - O(k) \leq f_k^{(3)}(H) \leq f_k^{(4)}(H)$. Finally, by scaling x and j by $(k-1)$, we define

$$c_H := \tfrac{1}{(k-1)^2} f_k^{(4)}(H) \quad = \quad \begin{array}{ll} \max & jx \\ \text{s.t.} & (t'-1)x + (d_{t'}-1)j \leq 1 \quad \text{for all } t' \in [t] \\ & x, j \geq 0 \end{array}$$

which is independent of k and depends only on the 1-graph H. As $\mathcal{E}_k^*(H) = f_k^{(4)}(H) \pm O(k)$, we finally attain $\mathcal{E}_k^*(H) = \big(c_H + o(1)\big)k^2$.

Now, although the program for c_H is not linear, it is clearly solvable in polynomial time. Further, notice that for $(x, j) = (\frac{1}{2(t-1)}, \frac{1}{2(d_1-1)})$, we have

$$(t'-1)x + (d_{t'}-1)j \leq \frac{1}{2} + \frac{1}{2} = 1,$$

for all $t' \in [t]$, so $c_H \geq \frac{1}{4(t-1)(d_1-1)}$. Additionally, only considering the constraints $(1-1)x + (d_1-1)j \leq 1$ and $(t-1)x + (d_t-1)j \leq 1$, we find that $x \leq \frac{1}{t-1}$ and $j \leq \frac{1}{d_1-1}$, so $c_H \leq \frac{1}{(t-1)(d_1-1)}$. □

4.4 Conclusion and Further Directions

In our study of the extremal function $\mathcal{E}_k(\mathcal{H})$, the largest open question is whether or not $\mathcal{E}_k(\mathcal{H}) = \mathcal{E}_k^*(\mathcal{H})$ when $\mathcal{H}$ is a family of simple, r-uniform graphs (see Conjecture 4.2.25); also very natural is the question of the behavior of $\mathcal{E}_k(C_4)$ (also discussed in the Introduction). Note for example that $\Omega(k^{4/3}) \leq \mathcal{E}_k(C_4) \leq O(k^{3/2})$ where the upper bound follows from Proposition 4.3.8 and the lower bound follows from the fact that $\mathrm{ex}(K_n, C_4) = \Theta(n^{3/2})$.

Several further questions follow naturally from our line of inquiry. For example:

Question 4.4.1. *What are the exact asymptotics of $\mathcal{E}_k(P_t)$?*

We note that, for a fixed t, $\mathcal{E}_k(P_t) = \Theta(k^2)$ where the upper bound follows from Corollary 4.2.24 or Proposition 4.3.8 and the lower bound follows from the fact that $\mathrm{ex}(K_n, P_t) = \frac{t-1}{2}n$. More specifically, what are the extremal graphs for $\mathcal{E}_k(P_t)$? Corollary 4.2.4 implies that there are extremal graphs for $\mathcal{E}_k(P_t)$ with diameter at most t. Gyárfás, Rousseau and Schelp [29] prove that if n is sufficiently large compared to t, then

$$\mathrm{ex}(K_{n,n}, P_t) = \begin{cases} \frac{t-1}{2}(2n - t + 1) & \text{for } t \text{ odd;} \\ \frac{t-2}{2}(2n - t + 2) & \text{for } t \text{ even.} \end{cases}$$

This implies that for all $t \geq 5$ and n sufficiently large, $\mathrm{ex}(K_{\sqrt{2}n}, P_t) < \mathrm{ex}(K_{n,n}, P_t)$, despite having the same number of edges, so most likely, the extremal graphs for $\mathcal{E}_k(P_t)$ look more

similar to cliques, as we showed was the case with P_3. However, $\mathrm{ex}(K_{\sqrt{2}n}, P_4) \approx \frac{3}{\sqrt{2}}n > 2n \approx \mathrm{ex}(K_{n,n}, P_4)$, so it may very likely be the case that the extremal graphs for $\mathcal{E}_k(P_4)$ are bipartite. As there is this discrepancy, it would be very interesting to just determine the extremal graphs for $\mathcal{E}_k(P_4)$ and why P_4 may behave differently from P_t for all other t.

Next, in regard to necklaces (see Theorem 4.3.10), we found that there is a multigraph G on $(\phi - o(1))k$ edges with $\mathrm{ex}(G, \mathcal{O}_2) < k$, but whenever G' is a multigraph with $\mathrm{ex}(G', \mathcal{O}_2) < k$ where each vertex has at most one loop, then $|E(G')| \leq \frac{3}{2}k < \phi k$. As such, it seems natural to ask about how $\mathcal{E}_k^*(H)$ changes if H is a non-uniform graph and the edges of different uniformities are weighted differently to reflect the fact that there are more possible edges of uniformity 2 in the host graph than there are of uniformity 1: one could more generally define $\mathrm{ex}(G, H) := \max\{\sum_{e \in F} w(|e|) : F \subset G, F \ H\text{-free}\}$, where w is an arbitrary weighting of the uniformities.

Question 4.4.2. *How do $\mathcal{E}_k(H)$ and $\mathcal{E}_k^*(H)$ vary with the weight w for non-uniform graphs?*

In fact, since the main obstacle to forcing non-uniform graphs appears to be the edges having irreconcilable "types," which suggests asking equivalent questions in the uniform case by artificially enforcing distinct edge-types on graphs that are already uniform.

Question 4.4.3. *Suppose H is a graph consisting of both red and blue edges. How many edges can a red-blue colored graph G have such that any k-edge subgraph contains a copy of H (with the correct colors)?*

Finally, recall that we originally defined $\mathcal{E}_k$ by deciding that a host graph being "best at forcing" meant optimizing specifically the its edge count, but one could just as easily ask this for any other monotone graph parameter P. That is, we could study $\mathcal{E}_{P,k}(\mathcal{H}) := \sup\{P(G) : \mathrm{ex}(G, \mathcal{H}) < k\}$. One particularly interesting example may be when $P = \chi$, the chromatic number. In this case, $\mathcal{E}_{\chi,k}(K_{1,t})$ and $\mathcal{E}_{\chi,k}(tK_2)$ are not trivial.

Question 4.4.4. *If $\mathcal{H}$ is a family of (multi)(hyper)graphs, what is $\mathcal{E}_{\chi,k}(\mathcal{H})$?*

To this end, we quickly note that as any graph G has $|E(G)| \geq \binom{\chi(G)}{2}$, Theorem 4.1.1 implies that if $\mathcal{H}$ is a family of simple, 2-uniform graphs with $\rho(\mathcal{H}) = \rho \geq 3$, then $\mathcal{E}_{\chi,k}(\mathcal{H}) = \sqrt{\left(2 + \frac{2}{\rho - 2} + o(1)\right)k}$; so again, it is most interesting to focus on families of bipartite graphs.

Chapter 5

Uniform-Weight Vectors of Bounded Rank

5.1 Introduction

The field of extremal combinatorics deals with the asymptotic study of how parameters grow over increasing classes of discrete structures. Recently (see for example [14]), there has been growing interest in the study of an extremal theory for matroids. This includes an extremal theory for *representable* matroids, whose ground set is the set of columns of some matrix (and independence is given by linear independence).

One standard method for generating *random* representable matroids, see e.g. [16], is as follows. Construct a matrix representation M by generating m randomly chosen columns of some fixed weight k and length n. Indeed, when $k = 2$ and the base field is $\mathbb{F}_2$, this gives the graphic matroid of the Erdős-Rényi random graph $G_{n,m}$ (of which M acts as the vertex-edge incidence matrix).

Our desire is to settle perhaps the most natural extremal question in this setting: how large can m be, upon fixing the "size" of such a representable matroid? It makes little sense to fix the number n of rows, as then one can take all $m = \binom{n}{k}(q-1)^k$ weight-k column vectors, and every possible matrix will just consist of a subset of these columns. So instead, we fix the rank.

Let us take a step back. For a matrix M over the finite field $\mathbb{F}_q$, we are considering the following question:

Question 5.1.1. *What is the maximum number of distinct columns M can have, if each column has weight k, that is k nonzero entries, and M has rank $\leq r$?*

We denote this value by $\mathrm{ex}_q(r, k)$. We can answer this question if r is large enough:

Theorem 5.1.2. *For all k, there is an R_k not depending on q such that for all $r \geq R_k$,*

$$\mathrm{ex}_q(r,k) = \begin{cases} \binom{r+1}{k} & q = 2 \text{ and } k \text{ even}, \\ \binom{r}{k}(q-1)^k & \text{otherwise.} \end{cases}$$

When $k = q = 2$, this tells us that graphs of graphic matroid rank $\leq r$ have $\leq \binom{r+1}{2}$ edges: we previously noted this in Theorem 4.2.8 of Section 4, where it was shown for *every* r (not just those sufficiently large). Furthermore, the case $q = 2$ was a question asked by Ahlswede, Aydinian and Khachatrian [2]. Khachatrian (according to [7]) and Kramer [36] conjectured the above structure, and the latter proved it when the number of *rows* of the matrix is $r+1$. Our result confirms their conjecture, but only once r is large enough.

The nature of this question does not change much after replacing "nonzero" with "non-β" for an arbitrary $\beta \in \mathbb{F}_q$ (see Section 5.3, and more specifically Theorem 5.3.4, an affine variant we will use to prove this result). This effectively answers both questions of this type over $\mathbb{F}_2$. However, for other fields $\mathbb{F}_q$, "weight k" and "k 1's" have different meanings, suggesting a complementary version of the original question:

Question 5.1.3. *What is the maximum number of distinct columns M can have, if each column has k zeros, and M has rank $\leq r$?*

Denoting this by $\overline{\mathrm{ex}}_q(r,k)$, we have a corresponding result:

Theorem 5.1.4. *Suppose $\mathbb{F}_q \neq \mathbb{F}_2$. For all k, there is an $\bar{R}_k = \bar{R}_k(q)$ such that for all $r \geq \bar{R}_k$,*

$$\overline{\mathrm{ex}}_q(r,k) = \binom{r}{k}(q-1)^{r-k}.$$

Furthermore, in the context of both Theorem 5.1.2 and Theorem 5.1.4, we will show that the only examples attaining the equality have exactly r nonzero rows (unless $k = 0$, see Corollary 5.2.3). This corresponds to the "uniqueness of the cliques" in Theorem 4.2.8 where additional isolated vertices correspond to additional rows of all 0's here.

In fact, a result akin to Theorem 5.1.2 holds in a far more general setting. Suppose $\mathbb{F}$ is an arbitrary field (not necessarily finite). Let $\mathbf{L} = (L_1, \ldots, L_s)$ be a collection of disjoint finite sets $L_i \subset \mathbb{F}^*$ of nonzero labels. Then, for each s-tuple $\mathbf{k} = (k_1, \ldots, k_s)$ of positive integers, an "$(\mathbf{L}, \mathbf{k})$-*vector*" is defined to be one with exactly k_i entries in L_i for each i, and the rest equal to 0. Thus, a binary vector of weight w is an $(\mathbf{L}, \mathbf{k})$-vector for $\mathbf{L} = (\{1\})$ and $\mathbf{k} = (w)$.

The corresponding question in this setting is thus:

Question 5.1.5. *What is the maximum number of distinct columns M can have, if each column is an $(\mathbf{L}, \mathbf{k})$-vector, and M has rank $\leq r$?*

We denote this value by $\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r, \mathbf{k})$.

We will prove the following theorem in Section 5.3:

Theorem 5.1.6. *For all $\mathbf{k} = (k_1, \ldots, k_s)$, there is an $R_{\mathbf{k}}$ such that for all $r \geq R_{\mathbf{k}}$,*

$$\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r, \mathbf{k}) = \begin{cases} \binom{r+1}{\mathbf{k}} & \forall i \in [s]\ L_i = \{\ell_i\} \text{ and } \sum \ell_i k_i = 0 \\ \mathbf{L}^{\mathbf{k}} \binom{r}{\mathbf{k}} & \text{otherwise,} \end{cases} \tag{5.1}$$

where by $\binom{r}{\mathbf{k}}$ we mean the multinomial coefficient $\binom{r}{k_1,\ldots,k_s,r-\sum_i k_i}$, and by $\mathbf{L}^{\mathbf{k}}$ we mean the product $\prod_{i\in[s]} |L_i|^{k_i}$. Moreover, any extremal matrix M has only $r+1$ or r nonzero rows respectively.

It is helpful to keep the case $s = 1$ in mind, so that $\mathbf{k} = (k_1)$ and $\mathbf{L} = (L_1)$. Here, writing $\mathbf{k}$ and $\mathbf{L}$ in place of k_1 and $|L_1|$ respectively, the pieces of notation $\mathbf{L}^{\mathbf{k}}$ and $\binom{r}{\mathbf{k}}$ agree with their usual meanings. In particular, Theorem 5.1.2 follows from Theorem 5.1.6 by taking $L_1 := \mathbb{F}_q^\times = \mathbb{F}_q \backslash \{0\}$, a single list consisting of all nonzero elements of $\mathbb{F}_q$. In this case, $|L_1| = 1$ if and only if $q = 2$ and $\ell_1 = 1$, so the clause "$\sum \ell_i k_i = 0$" says precisely that k_1 is even.

There is nonempty (albeit rather small) overlap between Theorem 5.1.6 and the main theorem of Ahlswede, Aydinian and Khachatrian [2]. They considered this question in the case $\mathbb{F} = \mathbb{R}, s = 1$, and $\mathbf{L} = (\{1\})$, i.e. binary vectors *over the reals* of weight k, but managed to solve this for *every* r. In particular, the equality given in (5.1) was shown to break down precisely once $r < 2k$. As with their question for $q = 2$, this leads us to ask how small R_k can be made in general—we will discuss this a little more in Section 5.5.

5.2 Preliminaries, Notation

We first obtain nontrivial bounds for all 3 questions, by generalizing the setup further still.

For an arbitrary set $S \subset \mathbb{Z}^s_{\geq 0}$ of possible weight vectors, we say a column vector is an "$(\mathbf{L}, S)$-*vector*" whenever it is an $(\mathbf{L}, \mathbf{k})$-vector for some $\mathbf{k} \in S$, and denote by $\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r, S)$ the maximum size of a collection of $(\mathbf{L}, S)$-vectors whose rank is $\leq r$. We can define $\mathrm{ex}_q(r, T), \overline{\mathrm{ex}}_q(r, T)$ correspondingly when T is just a subset of nonnegative integers, and specifically write $\mathrm{ex}_q(r, \leq k), \overline{\mathrm{ex}}_q(r, \leq k)$ as shorthand for $\mathrm{ex}_q(r, \{0, 1, \ldots, k\}), \overline{\mathrm{ex}}_q(r, \{0, 1, \ldots, k\})$ respectively.

We can form a poset structure $\preceq$ on the set of weight vectors $\mathbb{Z}^s_{\geq 0}$ of length s by saying $\mathbf{k}' \preceq \mathbf{k}$ if and only if $k'_i \leq k_i$ in every coordinate i. Say that $S \subset \mathbb{Z}^s_{\geq 0}$ is a *down-set* if $\mathbf{k}' \in S$ whenever $\mathbf{k} \in S$ and $\mathbf{k}' \preceq \mathbf{k}$.

Lemma 5.2.1. *For any rank r, field $\mathbb{F}_q$ and weight k:*

$$\overline{\mathrm{ex}}_q(r, \leq k) = \sum_{i \leq k} \binom{r}{i} (q-1)^{r-i}.$$

Also, for any field $\mathbb{F}$, down-set $S \subset \mathbb{Z}^s_{\geq 0}$, weight vector $\mathbf{k}$ and list vector $\mathbf{L}$,

$$\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r,S) = \sum_{\mathbf{k}' \in S} \binom{r}{\mathbf{k}'} \mathbf{L}^{\mathbf{k}'}, \text{ and hence for any } \mathbb{F}_q, \mathrm{ex}_q(r, \leq k) = \sum_{i \leq k} \binom{r}{i} (q-1)^i.$$

Proof. It suffices to show "$\leq$", since the corresponding lower bounds are all immediate from considering matrices with precisely r rows (with all columns of weight $\leq k$ in the first case, or all $(\mathbf{L}, S)$-vectors of length r in the second).

For the second bound, given a matrix M of rank r, let $\mathcal{C}$ be its columns and $W = \langle \mathcal{C} \rangle$ be its column space. Since the row rank of M is also r, there exists a subset I of r of its rows such that the projection $W \to W|_I$ is an isomorphism. In particular, it is injective, and restricts to an injection on the original vectors $\pi : \mathcal{C} \hookrightarrow \mathcal{C}|_I$. For any $x \in \mathcal{C}$ and $i \in [s]$, the number of L_i-entries in $\pi(x)$ is $\leq$ that of x. Hence, if x is an $(\mathbf{L}, \mathbf{k})$-vector, then $\pi(x)$ is an $(\mathbf{L}, \mathbf{k}')$-vector for some $\mathbf{k}' \preceq \mathbf{k}$, and hence an $(\mathbf{L}, S)$-vector as S is a down-set. The desired bound is then obtained by counting all $(\mathbf{L}, S)$-vectors in $\mathcal{C}|_I \simeq \mathbb{F}^r$.

The proof of the bound on $\overline{\mathrm{ex}}$ is identical, with "zero-entries" and "vectors with k' zeros" in place of "L_i-entries" and "$(\mathbf{L}, \mathbf{k}')$-vectors" respectively. □

Corollary 5.2.2. *For any $q, \mathbb{F}, r, k, \mathbf{k}$ and $\mathbf{L}$,*

$$\overline{\mathrm{ex}}_q(r,k) \leq \sum_{i \leq k} \binom{r}{i} (q-1)^{r-i},$$

$$\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r,\mathbf{k}) \leq \sum_{\mathbf{k}' \preceq \mathbf{k}} \binom{r}{\mathbf{k}'} \mathbf{L}^{\mathbf{k}'}, \text{ and } \mathrm{ex}_q(r,k) \leq \sum_{i \leq k} \binom{r}{i} (q-1)^i.$$

In particular, we obtain the $k = 0$ case of Theorem 5.1.4:

Corollary 5.2.3. $\overline{\mathrm{ex}}_q(r,0) = (q-1)^r$. *Furthermore, any matrix M with no zeros, of rank r, with $(q-1)^r$ distinct columns, has r rows $\mathbf{u}_1, \ldots, \mathbf{u}_r$ such that every row is a scalar multiple of some $\mathbf{u}_i$.*

Proof. Taking $k = 0$ in the 3rd equality of Corollary 5.2.2 establishes $\overline{\mathrm{ex}}_q(r,0) \leq (q-1)^r$.

For any matrix M attaining this equality, in the above proof, we see that $\mathcal{C}|_I$ consists of *all* column vectors with no zeros, that is, $\mathcal{C}|_I \simeq (\mathbb{F}_q^\times)^r$. Letting $\mathbf{u}_1, \ldots, \mathbf{u}_r$ denote the rows of M given by I, this says that for every $\mathbf{v} \in (\mathbb{F}_q^\times)^r$ there is some j such that $\mathbf{v} = \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{r,j} \end{pmatrix}$. Now, suppose there is another row u of M. Since $\mathrm{rank}(M) = r = \dim(\langle \mathcal{C}|_I \rangle)$, the $\{\mathbf{u}_i\}$ form a basis for the row space of M, so $\mathbf{u} = \sum \lambda_i \mathbf{u}_i$ for some scalars $\lambda_i \in \mathbb{F}_q$. As M has no zeros, $0 \neq \sum \lambda_i u_{i,j} = \langle \mathbf{x}, \mathbf{v} \rangle$ for *every* j, writing $\mathbf{x} := (\lambda_1, \ldots, \lambda_r)$.

Now, since $\mathbf{u} \neq 0$, $\mathbf{x} \neq 0$ so some $\lambda_j \neq 0$. Now take any $j' \in [r]\backslash\{j\}$. Consider the q^2 vectors of the form $\mathbf{v}(\alpha, \beta) := (1, \dots, 1, \alpha, 1, \dots, 1, \beta, 1, \dots, 1) \in \mathbb{F}_q^r$ with α, β in positions j, j' respectively. For each $\alpha \neq 0$, we know $\langle \mathbf{x}, \mathbf{v}(1, \alpha)\rangle \in \mathbb{F}_q^\times$, and they are distinct since $\lambda_j \neq 0$. It follows $\langle \mathbf{x}, \mathbf{v}(1, 0)\rangle = 0$. As $q \geq 3$, we similarly find another $\beta \in \mathbb{F}_q^\times\backslash\{1\}$, also with $\langle \mathbf{x}, \mathbf{v}(\beta, 0)\rangle = 0$ by the same logic. Subtracting these gives $0 = \langle \mathbf{x}, (1-\beta)\mathbf{e}_{j'}\rangle = (1-\beta)\lambda_{j'}$, hence $\lambda_{j'} = 0$.

Since $j' \neq j$ was arbitrary, $\mathbf{u} = \lambda_j \mathbf{u}_j$, which is what we were trying to prove. □

Remark 5.2.4. The final part of the argument showed that any vector over $\mathbb{F}_q$ ($q \geq 3$) of weight ≥ 2 is orthogonal to a nonzero number of vectors with no zeros. Later, Lemma 5.4.2 will count this number explicitly.

5.3 Weight-k Proofs

Our proofs will also establish an affine variant of Theorem 5.1.6 for technical reasons. To state it, define the *a-rank*, or *affine rank*, of a set of vectors to be the smallest r so that any subset of $r+1$ vectors yield an *a-dependence*, where by an a-dependence we mean a nontrivial linear dependence whose coefficients sum to 0 in $\mathbb{F}$.

Notation 5.3.1. *We denote by* $\mathrm{aex}_{\mathbf{L}}(r, \mathbf{k})$ *the maximum size of a collection of* $(\mathbf{L}, \mathbf{k})$ *vectors of a-rank* $\leq r$. *(As* $\mathbb{F}$ *will always be fixed, we drop the dependence in this notation.)*

Notice that, in general, the a-rank of a collection is at least the rank of the collection. On the other hand, the a-rank of the columns of a matrix M is the same as the rank of the matrix M with an additional row of all 1's added. Thus, we have

$$\mathrm{rank}(M) \leq a\text{-rank}(M) = \mathrm{rank}\begin{pmatrix} M \\ 1 \cdots 1 \end{pmatrix} \leq \mathrm{rank}(M) + 1. \tag{5.2}$$

Moreover,

Remark 5.3.2. Suppose every $L_i = \{\ell_i\}$ has only one element, and that $\sum \ell_i k_i \neq 0$. Then $\mathrm{aex}_{\mathbf{L}}(r, \mathbf{k}) = \mathrm{ex}_{\mathbf{L}}(r, \mathbf{k})$. Indeed, this time $(1, \dots, 1) \in \mathrm{rowspan}(M)$ for any matrix M whose columns are $(\mathbf{L}, \mathbf{k})$-vectors, and so its a-rank and rank coincide.

In a similar spirit, we will also make frequent use of the following standard lemma:

Lemma 5.3.3. *Let* $\lambda, \mu \in \mathbb{F}$ *be distinct. Then*

$$a\text{-}rank\begin{pmatrix} \lambda \cdots \lambda & \mu \\ Y & \mathbf{v} \end{pmatrix} = a\text{-}rank\begin{pmatrix} \lambda \cdots \lambda \\ Y \end{pmatrix} + 1,$$

and hence equals a*-rank*$(Y)+1$ *(provided* $\lambda \neq 0$*), for any vector* $\mathbf{v}$ *and matrix* Y *over* $\mathbb{F}$ *with the same number of rows.*

Proof. It suffices to show "$\geq$", the other direction being trivial.
Let $r = a\text{-rank}\begin{pmatrix} \lambda \cdots \lambda & \mu \\ Y & \mathbf{v} \end{pmatrix}$ and take any r column vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ of Y. By definition of r, there is an a-dependence among $\binom{\lambda}{\mathbf{v}_1}, \ldots, \binom{\lambda}{\mathbf{v}_r}, \binom{\mu}{\mathbf{v}}$. Since the coefficients sum to 0, and $\lambda \neq \mu$, it follows the coefficient of $\binom{\mu}{\mathbf{v}}$ is 0, so in fact we have an a-dependence among $\binom{\lambda}{\mathbf{v}_1}, \ldots, \binom{\lambda}{\mathbf{v}_r}$. Thus, $a\text{-rank}\binom{\lambda \cdots \lambda}{Y} \leq r-1$, as desired. □

Theorem 5.3.4. *For all k, there is a $Q_{\mathbf{k}}$ such that for all $r \geq Q_{\mathbf{k}}$,*

$$\text{aex}_{\mathbf{L}}(r, \mathbf{k}) = \begin{cases} \binom{r}{\mathbf{k}} & L = (\{\ell_1\}, \ldots, \{\ell_s\}) \\ \binom{r-1}{\mathbf{k}}\mathbf{L}^{\mathbf{k}} & \textit{otherwise.} \end{cases}$$

Moreover, any extremal collection must consist of vectors which are zero except in r common positions (respectively, $r-1$ common positions).

Proof. For the lower bound, we simply take "all vectors of the maximum possible length"- but we must be cautious whether the maximum possible length is r or $r-1$. First suppose $L_i = \{\ell_i\}$ for each i. Let M be the matrix whose columns are all $\binom{r}{\mathbf{k}}$ $(\mathbf{L}, \mathbf{k})$-vectors of length r. Then $\text{rank}(M) = r-1$ if $\sum \ell_i k_i = 0$, and r otherwise, but in both instances $a\text{-rank}(M) = r$ (see (5.2) and Remark 5.3.2, respectively).

Meanwhile, if $\exists i : |L_i| > 1$, then the matrix of all $\mathbf{L}^{\mathbf{k}}\binom{r-1}{\mathbf{k}}$ $(\mathbf{L}, \mathbf{k})$-vectors of length $r-1$ has rank $r-1$, and again by (5.2) has a-rank $\leq r$.

Write $\text{aex}^*_{\mathbf{L}}(r, \mathbf{k})$ for $\text{aex}_{\mathbf{L}}(r + \mathbf{1}_{\{\exists i : |L_i| > 1\}}, k)$. For the upper bound, we will prove $\text{aex}^*_{\mathbf{L}}(r, \mathbf{k}) \leq \mathbf{L}^{\mathbf{k}}\binom{r}{\mathbf{k}}$ for $r \geq Q_{\mathbf{k}}$.

To begin, we will show that for any $\mathbf{k}$ and for all $r \geq \|\mathbf{k}\| + 2$,

$$\text{aex}^*_{\mathbf{L}}(r, \mathbf{k}) \leq \text{aex}^*_{\mathbf{L}}(r-1, \mathbf{k}) + \sum_{i \in [s]} |L_i| \cdot \text{aex}^*_{\mathbf{L}}(r-1, \mathbf{k} - \mathbf{e}_i). \tag{5.3}$$

where $\mathbf{e}_i$ denotes the ith unit vector, so that $\mathbf{k} - \mathbf{e}_i = (k_1, \ldots, k_i - 1, \ldots, k_s)$. To see this, consider any matrix M of a-rank $\leq r$ whose columns are all $(\mathbf{L}, \mathbf{k})$-vectors. If all nonzero rows of M had all entries in $\bigcup L_i$, then M has only $\|\mathbf{k}\|$ nonzero rows and in particular $\leq \mathbf{L}^{\mathbf{k}}$ columns, independently of r. Plus, having already established the lower bound in the theorem, we know $\mathbf{L}^{\mathbf{k}} \leq \binom{r-2}{\mathbf{k}}\mathbf{L}^{\mathbf{k}} \leq \text{aex}^*_{\mathbf{L}}(r-1, \mathbf{k})$, only needing $r \geq 2 + \|\mathbf{k}\|$. So WLOG, assume that the first row of M contains both a 0 and an $\ell \in \bigcup L_i$.

Now let A_ℓ be the set of vectors with ℓ in row 1, for each $\ell \in \{0\} \cup \bigcup L_i$. Both $\bigcup_{\ell \neq 0} A_\ell$ and A_0 are nonempty by assumption. Define A'_ℓ to be the collection of vectors produced by removing the first coordinate from each vector in A_ℓ. By Lemma 5.3.3, A'_ℓ has a-rank $\leq r-1$ for every $\ell \in \{0\} \cup \bigcup L_i$. Hence $|A'_\ell| \leq \text{aex}_{\mathbf{L}}(r-1, \mathbf{k} - \mathbf{e}_i)$ whenever $\ell \in L_i$ while $|A'_0| \leq \text{aex}_{\mathbf{L}}(r-1, \mathbf{k})$. This establishes (5.3).

The inequality (5.3) would suffice to prove Theorem 5.3.4 if we could establish a family of base cases for the induction. We do not know how to do this directly, however. Instead we

define

$$\alpha_r^{\mathbf{k}} = \mathrm{aex}^*_{\mathbf{L}}(r, \mathbf{k}) - \mathbf{L}^{\mathbf{k}} \binom{r}{\mathbf{k}}$$

and consider the sequence $\{\alpha_r^{\mathbf{k}}\}_{r\in\mathbb{N}}$. Now, (5.3) gives that for $r \geq \|\mathbf{k}\| + 2$, $\alpha_r^{\mathbf{k}} \leq \alpha_{r-1}^{\mathbf{k}} + \sum_{i\in[s]} |L_i| \cdot \alpha_{r-1}^{\mathbf{k}-\mathbf{e}_i}$. By induction on $\|\mathbf{k}\|$, we then have for $Q'_{\mathbf{k}} := \max\{\|\mathbf{k}\|+2\} \cup \{Q_{\mathbf{k}-\mathbf{e}_i} : i \in [s]\}$ that

$$r - 1 \geq Q'_{\mathbf{k}} \implies \alpha_r^{\mathbf{k}} \leq \alpha_{r-1}^{\mathbf{k}}.$$

Observe that to prove Theorem 5.3.4, it suffices to show that for $r - 1 \geq Q'_{\mathbf{k}}$,

Claim 5.3.5.

$$\alpha_r^{\mathbf{k}} = \alpha_{r-1}^{\mathbf{k}} \implies (\alpha_r^{\mathbf{k}} = 0 \textit{ and any collection realizing } \mathrm{aex}^*_{\mathbf{L}}(r, \mathbf{k}) \textit{ must have support } r).$$

To prove Claim (5.3.5), let us suppose that $r, \mathbf{k}$ are such that $\alpha_r^{\mathbf{k}} = \alpha_{r-1}^{\mathbf{k}}$, and $r - 1 \geq Q'_{\mathbf{k}}$, so that

$$\mathrm{aex}^*_{\mathbf{L}}(r, \mathbf{k}) = \mathrm{aex}^*_{\mathbf{L}}(r-1, \mathbf{k}) + \sum_{i\in[s]} |L_i|^{k_i} \binom{r-1}{\mathbf{k}-\mathbf{e}_i}. \tag{5.4}$$

Recall that in the decomposition above, A_0 has size at most $\mathrm{aex}^*_{\mathbf{L}}(r-1, \mathbf{k})$. Thus for an extremal collection for $r, \mathbf{k}$ where (5.4) holds, we have that $\sum_{\ell\neq 0} |A_\ell| \geq \sum |L_i|^{k_i} \binom{r-1}{\mathbf{k}-\mathbf{e}_i}$. Moreover, by induction on $\|\mathbf{k}\|$, we have that the unique candidate for A'_ℓ of size $|L_i|^{\mathbf{k}-\mathbf{e}_i}\binom{r-1}{\mathbf{k}-\mathbf{e}_i}$ is a collection of vectors whose support has size $r-1$, for every $\ell \in L_i$.

In fact, these supports must be identical for every $\ell \in \cup L_i$. Indeed, suppose A'_ℓ contains a vector u such that $u(t) \in \cup L_i$, where t is outside the support of A'_j for some $j \in (\cup L_i)\backslash\ell$.

Then, using row t in Lemma 5.3.3 shows a-rank$(A'_j) \leq a$-rank$(A'_j|u) - 1 \leq r - 2$, a contradiction.

Furthermore, this means the support of A_0 must be contained in the support of A'_j (now equivalent for any $j \in \cup L_i$), establishing Claim (5.3.5), and thus also Theorem 5.3.4. Indeed, suppose A_0 contains a vector u such that $u(t) \in \cup L_i$, where t is outside the support of A'_j. This time we consider two cases:

Case 1: All vectors $v \in A_0$ satisfy $v(t) \in \cup L_i$, but A_0 is nonconstant on row t. Decomposing A_0 according to t-th entries, and applying Lemma 5.3.3 to each part we see $|A_0| \leq \sum |L_i| \cdot \mathrm{aex}^*_{\mathbf{L}}(r-1, \mathbf{k}-\mathbf{e}_i)$. So by induction on k, $|A_0| \leq \sum |L_i| \cdot \mathbf{L}^{\mathbf{k}-\mathbf{e}_i}\binom{r-1}{\mathbf{k}-\mathbf{e}_i} = O(r^{\|\mathbf{k}\|-1}) < \mathrm{aex}^*_{\mathbf{L}}(r-1, \mathbf{k})$ for r large enough, but this contradicts (5.4).

Case 2: All vectors $v \in A_0$ satisfy $v(t) = \ell \in L_i$. Deleting row t from A_0 then does not affect the a-rank, so in fact $|A_0| \leq \mathrm{aex}_{\mathbf{L}}(r-1, \mathbf{k}-\mathbf{e}_i) \leq O(r^{\|\mathbf{k}\|-1})$ by induction, again a contradiction.

Case 3: There is a vector $v \in A_0$ with $v(t) = 0$. In this case, two applications of Lemma 5.3.3, using rows t and 1 in turn, show a-rank$(M) \geq a$-rank$(u|v|A'_j) = a$-rank$(v|A'_j) + 1 = a$-rank$(A'_j) + 2 = r + 1$, a contradiction.

□

Remark 5.3.6. We can obtain an explicit bound on $Q_{\mathbf{k}}$ as follows. Claim (5.3.5) was sufficient to prove the theorem since $\{\alpha_r^{\mathbf{k}}\}$ is bounded below by 0. But in fact, by recalling $\operatorname{rank}(M) \le a\text{-}\operatorname{rank}(M)$ and applying Corollary 5.2.2,

$$\begin{aligned}
\alpha_{Q'_{\mathbf{k}}}^{\mathbf{k}} \le \operatorname{aex}^*_{\mathbf{L}}(Q'_{\mathbf{k}}, \mathbf{k}) \le \operatorname{ex}_{\mathbf{L}}(Q'_{\mathbf{k}} + 1, \mathbf{k}) &\le \sum_{\mathbf{k}' \preceq \mathbf{k}} \binom{Q'_{\mathbf{k}} + 1}{\mathbf{k}'} \mathbf{L}^{\mathbf{k}'} \\
&\le \binom{Q'_{\mathbf{k}} + 1}{\mathbf{k}} \mathbf{L}^{\mathbf{k}} \sum_{\mathbf{k}' \preceq \mathbf{k}} \left(\frac{\|\mathbf{k}'\|}{r}\right)^{\|\mathbf{k}-\mathbf{k}'\|} \\
&\le \binom{Q'_{\mathbf{k}} + 1}{\mathbf{k}} \mathbf{L}^{\mathbf{k}} \sum_{i=0}^{\|\mathbf{k}\|} \binom{\|\mathbf{k}\|}{i} \left(\frac{\|\mathbf{k}\|}{r}\right)^{i} \\
&= \binom{Q'_{\mathbf{k}} + 1}{\mathbf{k}} \mathbf{L}^{\mathbf{k}} \left(1 + \frac{\|\mathbf{k}\|}{r}\right)^{\|\mathbf{k}\|} \\
&\le O\big((Q'_{\mathbf{k}})^{\|\mathbf{k}\|}\big).
\end{aligned}$$

So the decreasing sequence $\{\alpha_r^{\mathbf{k}}\}$ stabilizes after $\le O\big((Q'_{\mathbf{k}})^{\|\mathbf{k}\|}\big)$ additional steps. Thus we can take $Q_{\mathbf{k}} := Q'_{\mathbf{k}} + O\big((Q'_{\mathbf{k}})^{\|\mathbf{k}\|}\big)$ in the theorem. This way, $Q_{\mathbf{k}}$ is bounded by $2^{O(\|\mathbf{k}\|^2)}$ as $\|\mathbf{k}\| \to \infty$.

We are now ready to prove Theorem 5.1.6, and make the transition from affine rank to usual rank.

Proof of Theorem 5.1.6. Note that if every $L_i = \{\ell_i\}$ and $\sum \ell_i k_i \neq 0$, the a-rank and rank coincide, and we are immediately done by Theorem 5.3.4.

So next suppose some $|L_i| > 1$. If we take the $\mathbf{L}^{\mathbf{k}}\binom{r}{\mathbf{k}}$ $(\mathbf{L}, \mathbf{k})$-vectors with some fixed support of size r, then the rank is exactly r. This gives the lower bound.

Now consider any collection of $(\mathbf{L}, \mathbf{k})$-vectors of rank at most r. By (5.2), the a-rank is $\le r + 1$, and Theorem 5.3.4 shows the size is at most $\mathbf{L}^{\mathbf{k}}\binom{r}{\mathbf{k}}$, along with the uniqueness of the equality case.

Lastly we consider the case where $\forall i L_i = \{\ell_i\}$ but $\sum \ell_i k_i = 0$.

For the lower bound, if we now take the $\mathbf{L}^{\mathbf{k}}\binom{r+1}{\mathbf{k}}$ $(\mathbf{L}, \mathbf{k})$-vectors with some fixed support of size $r + 1$, then the rank is at most r, since they all lie in the subspace $x \cdot (1, \dots, 1) = 0$.

For the upper bound, any collection of $(\mathbf{L}, \mathbf{k})$-vectors of rank at most r has a-rank at most $r + 1$, and we finish by Theorem 5.3.4 again, this time concluding they number $\le \mathbf{L}^{\mathbf{k}}\binom{r+1}{\mathbf{k}}$.

□

We may generalise Theorem 5.1.6 to an arbitrary set $S \subset \mathbb{Z}^s_{\ge 0}$ of possible weight vectors as follows. Recall that a vector is an $(\mathbf{L}, S)$-vector whenever it is an $(\mathbf{L}, \mathbf{k})$-vector for some

$\mathbf{k} \in S$, and that $\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r, S)$ is the maximum size of a collection of $(\mathbf{L}, S)$-vectors whose rank is $\leq r$.

Corollary 5.3.7. *For all S, there is an R_S such that for all $r \geq R_S$,*

$$\mathrm{ex}_{\mathbb{F},\mathbf{L}}(r, S) = \begin{cases} \sum_{\mathbf{k} \in S} \binom{r+1}{\mathbf{k}} & \forall i \in [s]\ L_i = \{\ell_i\} \text{ and } \forall \mathbf{k} \in S\ \sum \ell_i k_i = 0, \\ \sum_{\mathbf{k} \in S} \mathbf{L}^{\mathbf{k}} \binom{r}{\mathbf{k}} & \text{otherwise.} \end{cases}$$

Moreover, any extremal matrix M has only $r+1$ or r nonzero rows respectively.

This corollary is a huge generalization of Lemma 5.2.1. Although we need r to be sufficiently large (unlike previously), we now have an answer to all variants of questions of the form, "*How many columns with two 2's and a 1, or seven 4's, or of weight one, can a matrix over $\mathbb{F}_5$ of rank $\leq r$ have?*" In this instance, one would take $\mathbf{L} := (\{1\}, \{2\}, \{3\}, \{4\})$ and $S := \{(1,2,0,0), (0,0,0,7), (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1)\}$.

Proof. The lower bound arises from the collection of all $(\mathbf{L}, S)$-vectors with a given support of size $r+1$ (respectively, r) as before.

Take any collection $\mathcal{C}$ of $(\mathbf{L}, S)$-vectors with rank $\leq r$ and write $\mathcal{C}_{\mathbf{k}}$ for the set of $(\mathbf{L}, \mathbf{k})$-vectors in $\mathcal{C}$, for each $\mathbf{k} \in S$.

If $|L_i| > 1$ for some i, then $|\mathcal{C}_{\mathbf{k}}| \leq \mathbf{L}^{\mathbf{k}} \binom{r}{\mathbf{k}}$ for every $\mathbf{k} \in S$ by Theorem 5.1.6, and the desired upper bound on $|\mathcal{C}|$ follows. Furthermore, if equality holds, we know that $\mathcal{C}_{\mathbf{k}}$ has support r for every $\mathbf{k} \in S$, and furthermore these supports coincide for each k (otherwise any $v \in \mathcal{C}_{\mathbf{k}'}$ with support outside that of $\mathcal{C}_{\mathbf{k}}$ for some $\mathbf{k} \neq \mathbf{k}'$ increases the rank of $\mathcal{C}_{\mathbf{k}}$ from r to $\geq r+1$ by Lemma 5.3.3).

If every $L_i = \{\ell_i\}$ *and* $\sum \ell_i k_i' = 0$ for every $\mathbf{k}' \in S$, then $|\mathcal{C}| \leq \sum_{\mathbf{k}' \in S} \binom{r+1}{\mathbf{k}'}$. Supports having size $r+1$ and coinciding then both follow in the equality case exactly as before.

So we may assume that every $L_i = \{\ell_i\}$ but at least one $\mathbf{k} \in S$ has $\sum \ell_i k_i \neq 0$. Certainly, $|\mathcal{C}_{\mathbf{k}}| \leq \binom{r}{\mathbf{k}}$ whenever $\mathbf{k} \in S$ satisfies $\sum \ell_i k_i \neq 0$ as before. Here, we claim that in fact $\mathrm{rank}(\mathcal{C}_{\mathbf{k}'}) \leq r-1$ for every $\mathbf{k}' \in S$ which *does* satisfy $\sum \ell_i k_i' = 0$. Indeed, if $\mathbf{k} \in S$ has $\sum \ell_i k_i \neq 0$, and $\mathbf{v} \in \mathcal{C}_{\mathbf{k}}$ is chosen arbitrarily, then $r \geq \mathrm{rank}(\mathcal{C}_{\mathbf{k}'} \cup \{\mathbf{v}\}) = \mathrm{rank}(\mathcal{C}_{\mathbf{k}'}) + 1$, as $\mathbf{v}$ does not lie in the hyperplane $(1, 1, \ldots, 1)^{\perp}$ (whereas $\mathcal{C}_{\mathbf{k}'}$ does). It follows from Theorem 5.1.6 that $|\mathcal{C}_{\mathbf{k}'}| \leq \binom{r}{\mathbf{k}'}$ whenever $\sum \ell_i k_i' = 0$ too, giving the upper bound, and the equality case follows. □

5.4 k Zeros Proofs

We now proceed with the proof of Theorem 5.1.4. First, we establish a standard counting function:

Lemma 5.4.1. *For any sequence $u_1, u_2, \ldots$ of nonzero elements of $\mathbb{F}_q$, and any number n, the number of vectors $x \in (\mathbb{F}_q^\times)^n$ orthogonal to $(u_1, \ldots, u_n)$ is $a_n^{(0)} = \frac{1}{q}\big((q-1)^n + (-1)^n(q-1)\big)$.*

Proof. More generally, let $a_n^{(\beta)} := |S_n^\beta|$, where $S_n^\beta := \{x \in (\mathbb{F}_q^\times)^n : x_1u_1 + \cdots + x_nu_n = \beta\}$. Since $x \mapsto \beta x$ is a bijection $S_n^1 \to S_n^\beta$ for every $\beta \in \mathbb{F}_q^\times$, it follows $a_n^{(\beta)} = a_n^{(1)}$. Furthermore, $|S_{n+1}^\alpha| = \sum_{\beta \neq \alpha} |S_n^\beta|$ for any $\alpha \in \mathbb{F}_q$, since any vector in $\bigsqcup_{\beta\neq\alpha} S_n^\beta$ can be uniquely extended to a vector in S_n^α, since $v_n \neq 0$. Thus, we have the recursive relations for each $n \geq 0$:

$$\begin{aligned} a_{n+1}^{(0)} &= (q-1)a_n^{(1)}, \\ a_{n+1}^{(1)} &= (q-2)a_n^{(1)} + a_n^{(0)}. \end{aligned}$$

Since $a_0^{(\beta)} = \mathbf{1}_{\beta=0}$, the results $a_n^{(0)} = \frac{1}{q}\big((q-1)^n + (-1)^n(q-1)\big)$ and $a_n^{(1)} = \frac{1}{q}\big((q-1)^n + (-1)^{n+1}\big)$ follow by a trivial induction (or may be derived directly using generating functions). □

For a fixed r, we use X as shorthand for $\mathbb{F}_q^r$. Furthermore, for each $n \leq r$, we denote by $X^{\geq n}$ and $X^{=n}$, the sets of vectors of weight $\geq n$ and exactly n respectively. Immediately note that $|X^{=n}| = \binom{r}{n}(q-1)^n$ for every n.

Lemma 5.4.2. *Suppose $\mathbf{v} \in X^{\geq 2}$ has $i \geq 2$ non-zero entries, and W is its orthogonal complement $W := \mathbf{v}^\perp = \{\mathbf{x} \in X : \mathbf{x} \cdot \mathbf{v} = 0\}$. Then*

$$\begin{aligned} |X^{=r-k} \cap W| &= \frac{1}{q}\left(\binom{r}{k}(q-1)^{r-k} + (-1)^i(q-1)^{r-i-k+1}\sum_{s=0}^{k}(1-q)^s\binom{i}{s}\binom{r-i}{k-s}\right) \\ &\geq \frac{1}{q}\binom{r}{k}(q-1)^{r-k}\left(1 - \frac{1}{(q-1)}\right) \text{ for } r \text{ sufficiently large.} \end{aligned}$$

Proof. WLOG, $\mathbf{v} = (v_1, \ldots, v_i, 0, \ldots, 0)$ where $v_1, \ldots, v_i$ are all non-zero.

For each $S \in \binom{[r]}{k}$, let $W_S := \{\mathbf{x} \in X^{=r-k} \cap W : \{j \in [r] : x_j = 0\} = S\}$, so we may decompose W as $\bigcup_{s=0}^k \bigcup_{|S\cap[i]|=s} W_S$,

We claim that, if $|S \cap [i]| = s$, then $|W_S| = \frac{1}{q}\left((q-1)^{r-k} + (-1)^{i+s}(q-1)^{r-i-k+s+1}\right)$. Since there are $\binom{i}{s}\binom{r-i}{k-s}$ such $S \in \binom{[r]}{k}$ with $|S \cap [i]| = s$, the above decomposition gives the result.

To see the claim, note

$$\mathbf{x} \in W_S \Leftrightarrow \begin{cases} x_j = 0 & \forall j \in S \\ x_j \neq 0 & \forall j \in [r]\backslash S \\ \sum\limits_{j\in[r]\backslash S} x_jv_j = 0 & \end{cases} \Leftrightarrow \begin{cases} x_j = 0 & \forall j \in S \\ x_j \neq 0 & \forall j \in [r]\backslash(S \cup [i]) \\ x_j \neq 0 & \forall j \in [i]\backslash S \end{cases} \wedge \sum_{j\in[i]\backslash S} x_jv_j = 0.$$

Applying the lemma to $(u_1, \ldots, u_n) := \mathrm{proj}_{[i]\backslash S}(\mathbf{v})$ and noting $n = i - s$, we see there are $\frac{1}{q}\big((q-1)^{i-s} + (-1)^{i-s}(q-1)\big)$ ways to choose the entries of $\mathbf{x}$ in $S \cap [i]$. Furthermore, there

are $(q-1)^{|[r]\setminus(S\cup[i])|} = (q-1)^{r-k-i+s}$ ways to choose the entries of $\mathbf{x}$ in $[r]\setminus(S\cup[i])$, and so there are $\frac{1}{q}\left((q-1)^{r-k}+(-1)^{i+s}(q-1)^{r-i-k+s+1}\right)$ such $\mathbf{x}$ in total.

We now proceed to prove the claimed inequality. Let $a_s := \binom{i}{s}\binom{r-i}{k-s}(q-1)^s$, for each $0 \le s \le k$.

Note that, for $s < i$, $\frac{a_s}{a_{s+1}} = \frac{(s+1)(r-k-i+s+1)}{(i-s)(k-s)(q-1)}$ is an increasing function in s (and for $s > i$ $a_s = 0$ anyway). Hence the sequence $\{a_s\}$ is *unimodal*, i.e. consists of a (possibly empty) monotonically increasing subsequence followed by a decreasing subsequence. In particular, the alternating sum $\sum_{s=0}^{k}(-1)^s a_s$ is bounded above by $\max_s\{a_s\}$. Let us fix the s attaining this maximum.

Now,

$$\frac{\binom{r}{k}(q-1)^{i-1}}{a_s} = (q-1)^{i-1-s}\frac{\binom{r}{k}}{\binom{i}{s}\binom{r-i}{k-s}} \ge q-1$$

provided $i \ge s+2$, since the denominator is a single term in the identity $\sum_{s'}\binom{i}{s'}\binom{r-i}{k-s'} = \binom{r}{k}$. Else, $i \in \{s, s+1\}$. We check the lower bound still holds here:

When $i = s+1$, the above is $\frac{\binom{r}{k}}{i\binom{r-i}{k-i+1}} \ge \frac{\binom{r}{k}}{i\binom{r-i}{k-i}} = \frac{\binom{r}{i}}{i\binom{k}{i}} \ge \frac{(r-1)^2}{i(k-1)^2}$ (using $i \ge 2$ and e.g. $r \ge 2k$).

Similarly, if $i = s$, the above is $\frac{\binom{r}{k}}{(q-1)\binom{r-i}{k-i}} = \frac{\binom{r}{i}}{(q-1)\binom{k}{i}} \ge \frac{(r-1)^2}{(q-1)(k-1)^2}$.

So these are both still $\ge q-1$, assuming $r \ge \max\{q^{1/2}k^{3/2}, qk\}$. In summary,

$$\begin{aligned}\frac{1}{q}(q-1)^{r-i-k+1}\left(\binom{r}{k}(q-1)^{i-1}+\sum_{s'=0}^{k}(-1)^{s'+i}a_{s'}\right) &\ge \frac{1}{q}(q-1)^{r-i-k+1}\left(\binom{r}{k}(q-1)^{i-1}-a_s\right)\\ &\ge \frac{1}{q}\binom{r}{k}(q-1)^{r-k}\left(1-\frac{1}{q-1}\right).\end{aligned}$$

□

We are now in a position to prove the nonzero case of Theorem 5.1.4. This time, the extremal matrices cannot have *any* duplicate rows, nor scalings thereof.

Theorem 5.4.3. *Let $k \ge 1$. Then $\overline{\mathrm{ex}}_q(r,k) = \binom{r}{k}\cdot(q-1)^{r-k}$, provided $r \ge \max\{3q^2k, q^{1/2}k^{3/2}\}$. Furthermore, the unique extremal example is a matrix M consisting of only r rows and all possible columns.*

Proof. We may assume $\mathrm{rank}(M) = r$, and that all rows are distinct. Let Y denote the set of columns of M, with span $\langle Y\rangle = V$. If r' denotes the number of rows of M, then $V \le \mathbb{F}_q^{r'}$ is a subspace of dimension r.

For each $j \in [r']$, we have $V_j' := \{\mathbf{y} \in \mathbb{F}_q^{r'} : y_j = 0\}$ is codimension-1 in $\mathbb{F}_q^{r'}$, and hence $V_j := V_j' \cap V$ is codimension-≤ 1 in V. Whenever $\dim(V_j) = r-1$, we say that row j is *nontrivial*, and observe $V_j = \mathbf{v}_j^\perp$ for some $\mathbf{v}_j \in V$ (note that $\mathbf{e}_j$ is not necessarily in V). Otherwise, $V_j = V$, and we say row j is *trivial*. (In fact, every trivial row of M is necessarily all zeros, but we will not need this for the argument.)

We have by assumption that every $\mathbf{y} \in Y$ is in *exactly* k of the $\{V_j\}$ counting multiplicities, and hence in κ of the $\{V_j : j \text{ nontrivial}\}$, where $\kappa := k - |\{j \text{ trivial}\}|$. Furthermore, $\bigcap_j V_j \leq \bigcap_j V_j' = \{0\}$. Taking orthogonal complements, $\langle\{\mathbf{v}_j : j \text{ nontrivial}\}\rangle = V$, and hence they contain a basis $B \subset \{\mathbf{v}_j\}$.

Let $T : V \to \mathbb{F}_q^r = X$ map this basis B to the standard basis $\{\mathbf{e}_1, \dots, \mathbf{e}_r\} = X^{=1}$ of X, extended linearly to an isomorphism. The assumption certainly tells us for every $\mathbf{y} \in Y$ that $T(\mathbf{y})$ is in exactly k of the mapped subspaces $\mathcal{F} := \{T(V_j) : j \text{ nontrivial}\}$ (viewed as a multiset). In particular, $\mathcal{F}$ contains every coordinate subspace $\mathbf{e}_\ell^\perp$ *at least once* (as $\mathbf{e}_\ell^\perp$ is $T(V_j)$ for some $\mathbf{v}_j \in B$). Every $T(\mathbf{y})$ is in $\leq \kappa$ of these coordinate subspaces, and hence has $\leq \kappa$ zeros. Deduce $T(Y) \subset X^{\geq r-\kappa}$, so we immediately obtain $|Y| = |T(Y)| \leq \binom{r}{\kappa}(q-1)^{r-\kappa} + \binom{r}{\kappa-1}(q-1)^{r-\kappa+1} + \cdots + (q-1)^r$. Of course, that was something we already established in Corollary 5.2.2, but we will need this setup to help remove the trailing terms.

Suppose first that there is some $W \in \mathcal{F}$ which is not a coordinate hyperplane. We will show that $|Y|$ is too small in this case. Now, by dimension counting, $W^\perp = \langle\mathbf{v}\rangle$ for some $\mathbf{v} \in X$. Plus, as W is not a coordinate hyperplane, $\mathbf{v}$ has ≥ 2 non-zero entries. Thus, the previous lemma shows there are many vectors of weight $r-\kappa$ in W.

In fact, $T(Y) \subset (X^{=r-\kappa} \backslash W) \cup (X^{\geq r-\kappa+1})$, since all vectors in $X^{\leq r-\kappa-1} \cup (X^{=r-\kappa} \cap W)$ are in $\geq \kappa+1$ spaces in $\mathcal{F}$. Also note that $|X^{\geq r-\kappa+1}| \leq 2|X^{=r-\kappa+1}| = 2\binom{r}{\kappa-1}(q-1)^{r-\kappa+1}$ provided $r \geq 2q\kappa$. Putting these together with Lemma 5.4.2,

$$\begin{aligned}
|Y| = |T(Y)| &\leq |X^{=r-\kappa}| - |X^{=r-\kappa} \cap W| + |X^{\geq r-\kappa+1}| \\
&\leq \binom{r}{\kappa}(q-1)^{r-\kappa} - \frac{1}{q}\binom{r}{\kappa}(q-1)^{r-\kappa}\left(1 - \frac{1}{(q-1)}\right) + 2\binom{r}{\kappa-1}(q-1)^{r-\kappa+1} \\
&< \binom{r}{\kappa}(q-1)^{r-\kappa}, \text{ if } r \geq 3q^2\kappa.
\end{aligned}$$

As such, we may assume every $T(V_j) \in \mathcal{F}$ is some coordinate hyperplane $\mathbf{e}_\ell^\perp$. However, we still are not yet sure that the original subspaces $\{V_j\}$ were distinct (in the way that the $\{V_j'\}$ are): there may be collisions upon intersection with V.

For each $\mathbf{x} \in \mathbb{F}_q^r$, we denote by $Z_\mathbf{x}$ its zero-set $\{\ell : x_\ell = 0\}$. Also, letting $w(\mathbf{x}) := |\{W \in \mathcal{F} : \mathbf{x} \in W\}|$ (counting multiplicities), we see that every $\mathbf{x} \in T(Y)$ has $w(\mathbf{x}) = \kappa$. Form a poset structure on $X = \mathbb{F}_q^r$ by $\mathbf{x} \prec \mathbf{y} \Leftrightarrow Z_\mathbf{x} \supsetneq Z_\mathbf{y}$: thus, $(X, \preceq)$ looks like a blowup of the Boolean lattice where each vector of weight n has been blown up $(q-1)^n$ times. Also, since $\mathcal{F}$ contains each coordinate subspace at least once, w is a strictly increasing function on $(X, \preceq)$, and hence $T(Y)$ forms an antichain.

This satisfies a LYM-type inequality (see e.g. [10] for an exposition we will mimic here): for any arbitrary $A \subset X$, write $A^{=i} := A \cap X^{=i}$ for each $i \leq r$. Then a random maximal chain C in X satisfies $\mathbb{E}[|C \cap A|] = \sum_{i \leq r} \frac{|A^{=i}|}{|X^{=i}|}$ by symmetry. For the antichain $A := T(Y)$, deduce this is ≤ 1. Furthermore, with $r \geq qk \geq q\kappa$, we have $|X^{=r-\kappa}| > |X^{=r-\kappa+1}| > \cdots > |X^{=r}|$, and hence

$$1 \geq \sum_{i \leq r} \frac{|A^{=i}|}{|X^{=i}|} = \sum_{r-\kappa \leq i \leq r} \frac{|A^{=i}|}{|X^{=i}|} \geq \sum_{r-\kappa \leq i \leq r} \frac{|A^{=i}|}{|X^{=r-\kappa}|} = \frac{|A|}{|X^{=r-\kappa}|},$$

so $|Y| = |A| \leq |X^{=r-\kappa}| = \binom{r}{\kappa}(q-1)^{r-\kappa} \leq \binom{r}{k}(q-1)^{r-k}$ is immediate. In the equality case, $k = \kappa$, and all rows were nontrivial. Also, every $\frac{|A^{=i}|}{|X^{=i}|} = \frac{|A^{=i}|}{|X^{=r-k}|}$ for $i > r-k$, hence they are all 0, from which it follows $T(Y) = A = X^{=r-k}$.

Deduce $\mathcal{F} = \{\mathbf{e}_1^\perp, \ldots, \mathbf{e}_r^\perp\}$ with no repeated subspaces, so $r' = r$ and M only had r rows originally. □

5.5 Concluding Remarks and Further Questions

In light of Theorems 5.1.2 and 5.1.4, one may naively hope that $\overline{\mathrm{ex}}_q(r,k) = \mathrm{ex}_q(r, r-k)$ for some reasonable values of r, k and q, but this is very far from being true, so there is a limit to how small we can make R_k and $\bar{R}_k$. Indeed, $\overline{\mathrm{ex}}_q(r,k)$ is an increasing function of k (for fixed r and q), since adding a row of zeros does not increase the rank of a matrix. Plus, while adding rows of all 1's might increase the rank, it does not increase the a-rank, so $\mathrm{aex}_{\mathbf{L}}(r, \mathbf{k})$ is also an increasing function of $\mathbf{k}$.

Even more strikingly, $\mathrm{ex}_q(r,k) = \overline{\mathrm{ex}}_q(r,k) = 0$ for negative k, whereas $\mathrm{ex}_q(r,k), \overline{\mathrm{ex}}_q(r,k)$ can be defined for $k > r$ and are clearly positive: in fact, $\mathrm{ex}_q(r, (q-1)q^{r-1}) = \overline{\mathrm{ex}}_q(r, q^{r-1}) = q^r - 1$. This is clearly the most possible for *any* k, since an $\mathbb{F}_q$-vector space of dimension r only has q^r distinct elements in total, including 0.

The matrix M attaining the above is simply the *dual Hamming code* [30] of length q^r, as noted in the concluding section of Ahlswede, Aydinian and Khachatrian [2] (and in fact, was shown to be essentially the unique such matrix up to repetition by Bonisoli [15]). Explicitly, we list all q^r vectors $\mathbb{F}_q^r = \{\mathbf{v}_1, \ldots, \mathbf{v}_{q^r}\}$ as the *rows* of a matrix A, then let the columns of M consist of all nonzero vectors in the column space of $A = (\mathbf{u}_1 | \ldots | \mathbf{u}_r)$, so $\mathrm{rank}(M) = r$. Now, the i-th entry of a column $\sum \lambda_j \mathbf{u}_j$ of M is zero if and only if $\mathbf{v}_i$ is in the hyperplane $\{\sum_j \lambda_j x_j = 0\}$. This is true for exactly q^{r-1} such vectors $\mathbf{v}_i \in \mathbb{F}_q^r$, and hence every column of M has weight $(q-1)q^{r-1}$.

So, we know these theorems cannot be extended arbitrarily. But we can still ask about the threshold functions:

Question 5.5.1. *How small can R_k and $\bar{R}_k$ be made in Theorems 5.1.2 and 5.1.4?*

Theorem 5.1.4 was established directly, obtaining the result for $\bar{R}_k = O(k^{3/2})$. In sharp

contrast, the proof of Theorem 5.1.2 used an induction for which we were unable to directly establish a base case, and is only known for $R_k = 2^{O(k^2)}$. However, the former grows with q, whereas the latter does not.

In fact, $\bar{R}_k$ *can't* be made independent of q. Once $r < qk$, we note that $\binom{r}{k}(q-1)^{r-k} < \binom{r}{k-1}(q-1)^{r-k+1}$, so (for example) our usual example is beaten by a matrix consisting of all co-weight $k-1$ vectors of length r, and then appending a row of all 0's. Perhaps it is still true that $\overline{\mathrm{ex}}_q(r,k) = \binom{r}{k'}(q-1)^{r-k'}$ for some $k' < k$, using matrices with lots of empty rows. Similar logic shows that R_k can't be made smaller than $\frac{q}{q-1}k \leq 2k$. Yet, it is still plausible that e.g. $\mathrm{ex}_2(2k,k) = \binom{2k}{k}$ for every odd k.

Furthermore, we wonder whether Theorem 5.1.4 can be generalized in a similar fashion to Theorem 5.1.6. This leads to questions that lie strictly between the original two, the simplest instance of which is the following:

Question 5.5.2. *Does every rank-r matrix over $\mathbb{F}_4$ have $\leq \binom{r}{k} \cdot 2^r$ columns with exactly k entries either 0 or 1 (for all r sufficiently large)?*

Index

Bibliography

[1] H. Abbott, D. Hanson, and N. Sauer. Intersection theorems for systems of sets. *Journal of Combinatorial Theory, Series A*, 12(3):381–389, 1972.

[2] R. Ahlswede, H. Aydinian, and L. Khachatrian. Maximum number of constant weight vertices of the unit n-cube contained in a k-dimensional subspace. *Combinatorica*, 23(1):5–22, 2003.

[3] M. Ajtai, J. Komlós, and E. Szemerédi. First occurrence of hamilton cycles in random graphs, cycles in graphs (burnaby, bc, 1982), 173–178. *Annals of DM*, 27.

[4] N. Alon. On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel Journal of Mathematics*, 38(1):116–130, 1981.

[5] N. Alon. Bipartite subgraphs. *Combinatorica*, 16(3):301–311, 1996.

[6] N. Alon and E. Halperin. Bipartite subgraphs of integer weighted graphs. *Discrete Mathematics*, 181(1-3):19–29, 1998.

[7] A. E. Ashikhmin, G. D. Cohen, M. Krivelevich, and S. N. Litsyn. Bounds on distance distributions in codes of known size. *IEEE transactions on information theory*, 51(1):250–258, 2005.

[8] T. Bohman, A. Frieze, M. Krivelevich, P.-S. Loh, and B. Sudakov. Ramsey games with giants. *Random Structures & Algorithms*, 38(1-2):1–32, 2011.

[9] B. Bollobás. The evolution of sparse graphs, in graph theory and combinatorics proceedings. In *Cambridge Combinatorial Conference in Honour of Paul Erdos*, pages 335–357, 1984.

[10] B. Bollobás. *Combinatorics: set systems, hypergraphs, families of vectors, and combinatorial probability*. Cambridge University Press, 1986.

[11] B. Bollobás. *Modern graph theory*, volume 184. Springer Science & Business Media, 2013.

[12] B. Bollobás and A. M. Frieze. On matchings and hamiltonian cycles in random graphs. Technical report, Carnegie-Mellon Univesity, Pittsburgh PA Management Sciences Research Group, 1983.

[13] B. Bollobás and A. D. Scott. Better bounds for max cut. *Contemporary combinatorics*, 10:185–246, 2002.

[14] J. Bonin. An introduction to extremal matroid theory with an emphasis on the geometric perspective (course notes). *Universitat Politecnica de Catalunya, Barcelona*, 2003.

[15] A. Bonisoli. Every equidistant linear code is a sequence of dual hamming codes. *Ars Combinatoria*, 18:181–186, 1984.

[16] C. Cooper, A. Frieze, and W. Pegden. Minors of a random binary matroid. *arXiv preprint arXiv:1612.02084*, 2016.

[17] P. Erdős. On extremal problems of graphs and generalized graphs. *Israel Journal of Mathematics*, 2(3):183–190, 1964.

[18] P. Erdős and R. Rado. Intersection theorems for systems of sets. *Journal of the London Mathematical Society*, 1(1):85–90, 1960.

[19] P. Erdős and A. Rényi. On the strength of connectedness of a random graph. *Acta Mathematica Academiae Scientiarum Hungarica*, 12(1-2):261–267, 1964.

[20] P. Erdos and A. H. Stone. On the structure of linear graphs. *Bull. Amer. Math. Soc*, 52(1087-1091):1, 1946.

[21] P. Erds and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5:17–61, 1960.

[22] W. Feller. The fundamental limit theorems in probability. *Bull. Amer. Math. Soc.*, 51:800–832, 1945.

[23] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 2008.

[24] S. Ferneyhough, R. Haas, D. Hanson, and G. MacGillivray. Star forests, dominating sets and ramsey-type problems. *Discrete mathematics*, 245(1-3):255–262, 2002.

[25] E. Friedgut and J. Kahn. On the number of copies of one hypergraph in another. *Israel Journal of Mathematics*, 105(1):251–256, 1998.

[26] A. Frieze and M. Karoński. *Introduction to random graphs*. Cambridge University Press, 2015.

[27] A. M. Frieze. An algorithm for finding hamilton cycles in random directed graphs. *Journal of Algorithms*, 9(2):181–204, 1988.

[28] Z. Füredi and M. Simonovits. The history of degenerate (bipartite) extremal graph problems. In *Erdős Centennial*, pages 169–264. Springer, 2013.

[29] A. Gyárfás, C. C. Rousseau, and R. H. Schelp. An extremal problem for paths in bipartite graphs. *Journal of graph theory*, 8(1):83–95, 1984.

[30] R. W. Hamming. Error detecting and error correcting codes. *Bell Labs Technical Journal*, 29(2):147–160, 1950.

[31] S. Janson, T. Luczak, and A. Rucinski. *Random graphs*, volume 45. John Wiley & Sons, 2011.

[32] P. Keevash. Hypergraph Turán problems.

[33] F. Knox, D. Kühn, and D. Osthus. Edge-disjoint hamilton cycles in random graphs. *Random Structures & Algorithms*, 46(3):397–445, 2015.

[34] J. Komlós and E. Szemerédi. Limit distribution for the existence of hamiltonian cycles in a random graph. *Discrete Mathematics*, 43(1):55–63, 1983.

[35] A. Korshunov. Solution of a problem of erdos and rényi on hamilton cycles in non-oriented graphs. In *Soviet Math. Dokl*, volume 17, pages 760–764, 1976.

[36] J. B. Kramer. On the most weight w vectors in a dimension k binary code. *The Electronic Journal of Combinatorics*, 17(1):142, 2010.

[37] M. Krivelevich, E. Lubetzky, and B. Sudakov. Hamiltonicity thresholds in achlioptas processes. *Random Structures & Algorithms*, 37(1):1–24, 2010.

[38] M. Krivelevich and W. Samotij. Optimal packings of hamilton cycles in sparse random graphs. *SIAM Journal on Discrete Mathematics*, 26(3):964–982, 2012.

[39] C. Lee, B. Sudakov, and D. Vilenchik. Getting a directed hamilton cycle two times faster. *Combinatorics, Probability and Computing*, 21(5):773–801, 2012.

[40] L. Lovász and M. D. Plummer. *Matching theory*, volume 367. American Mathematical Soc., 2009.

[41] C. McDiarmid. Clutter percolation and random graphs. In *Combinatorial Optimization II*, pages 17–25. Springer, 1980.

[42] L. Pósa. Hamiltonian circuits in random graphs. *Discrete Mathematics*, 14(4):359–364, 1976.

[43] P. Turán. On an external problem in graph theory. *Mat. Fiz. Lapok*, 48:436–452, 1941.

[44] V. Vizing. A bound on the external stability number of a graph. In *Dokl. Akad. Nauk SSSR*, volume 164, pages 729–731, 1965.

[45] D. West. *Introduction to graph theory*. Prentice Hall, Inc., Upper Saddle River, NJ, 1996.

[46] K. Zarankiewicz. Problem p101. In *Colloq. Math*, volume 2, page 5, 1951.