UNIVERSITY OF CALIFORNIA SAN DIEGO

# Topics in Random Graph Theory and Ramsey Theory

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Mathematics

by

Emily Zhu

Committee in charge:

Professor Lutz Warnke, Chair
Professor Frederick Manners, Co-Chair
Professor Ioana Dumitriu
Professor Daniel Kane

2023

The Dissertation of Emily Zhu is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2023

## DEDICATION

The author, Emily Zhu, died before completing this thesis. It is therefore dedicated to her memory.

TABLE OF CONTENTS

LIST OF FIGURES

# PREFACE

Emily Zhu began her PhD at the University of California San Diego in 2019. Tragically, she died suddenly in the summer of 2023. Her loss – as a talented researcher, but more importantly as a good friend – is felt sorely by all those here who knew her.

At the time of her death, Emily already had an impressive body of completed academic work. She co-authored three papers or preprints, with different co-authors, two of which had already been published in reputed journals (SIAM Journal on Discrete Mathematics, Proceedings of the American Mathematical Society) and the third of which has been submitted to a top journal in the field. Together, these constitute a masterful body of work in the area of combinatorics, and specifically in the areas of Ramsey theory and random graph theory.

Her co-advisors (Frederick Manners and Lutz Warnke) have compiled this dissertation from those three completed works. We note that Emily had a number of other promising research projects in progress, which might otherwise have formed part of her dissertation, but these do not form part of the work being considered here.

The work contained in this dissertation solves well-known and previously unsolved problems of general interest in combinatorics, and involves both innovative ideas and mastery of existing techniques. While it does not do full justice to her accomplishments as a mathematician, we hope that this compilation gives a partial testament to what she had achieved and would have gone on to achieve.

ACKNOWLEDGEMENTS

We do not know what personal thanks Emily would have written here. She would surely have wished to thank her family, her peers and others for their support.

On their own account, the compilers of this dissertation would like to thank her family for suggesting and pursuing the idea of completing this dissertation posthumously, and everyone at the University of California San Diego who assisted this process.

The following standard acknowledgements of co-authorship and use of published material are included to meet the UCSD specifications on dissertations.

Chapter 1 is a reprint, in full, of the material of the paper *A note on the Erdős–Hajnal hypergraph Ramsey problem*, which was published in *Proceedings of the American Mathematics Society*, volume 150(9), 2022, pp. 3675–3685. This paper was co-authored by the dissertation author together with Dhruv Mubayi and Andrew Suk.

Chapter 2, in full, is a reprint of the material as it appears in the paper *On multicolor Ramsey numbers of triple system paths of length* 3. This paper was published in *SIAM Journal on Discrete Mathematics*, volume 37(3), 2023, pp. 1419–1435. It was co-authored by the dissertation author together with Tom Bohman.

Chapter 3 is a reprint, in full, of the material as it appears in the preprint *Isomorphisms between random graphs*. This preprint was co-authored by the dissertation author together with Erlang Surya and Lutz Warnke. It was previously submitted for publication.

# VITA

| | |
|---|---|
| 2019 | Bachelor of Science / Master of Science in Mathematical Sciences, Carnegie Mellon University |

ABSTRACT OF THE DISSERTATION

Topics in Random Graph Theory and Ramsey Theory

by

Emily Zhu

Doctor of Philosophy in Mathematics

University of California San Diego, 2023

Professor Lutz Warnke, Chair
Professor Frederick Manners, Co-Chair

We present three separate chapters covering distinct results in combinatorics; more specifically, Ramsey theory and random graph theory. The first two chapters improve bounds on certain hypergraph Ramsey numbers. The third chapter determines sharp thresholds for problems related to the largest sub-structure that two random graphs have in common.

# Introduction

This dissertation discusses two problems in the area of hypergraph Ramsey theory, and one set of problems in the area of random graph theory.

The common theme behind these problems is finding a structured piece of an unstructured object. The *hypergraph Ramsey problem* asks, in general, the following: if all $r$-subsets of $\{1,\ldots,n\}$ are assigned one of $k$ colors arbitrarily, can we be sure to find some particular arrangement of $r$-subsets which all have the same color? The answer "yes, if $n$ is large enough" has been known since the 1930's, but understanding *how* large $n$ must be has been a central problem in combinatorics ever since. These are fundamental problems with applications both within and outside combinatorics.

In Chapter 1, the monochromatic substructure we wish to find is one copy of a particular 3-uniform hypergraph $\mathscr{L}$ (the "loose path") or $\mathscr{M}$ (the "messy path"), where $r = 3$. In this case $n$ may be taken to be linear in the number of colors $k$, and we show that the precise bounds $n < 1.54k$ and $n < 1.6k$ are sufficient.

In Chapter 2, we consider a more classical case, an instance of the notorious *Erdős–Hajnal problem*: using just two colors red and blue, we wish to find either a red $s$-clique or a blue $t$-clique. In some cases the correct rough growth rate of $n$ is known: we show that in the case $s = r+1$, $t = r-1$, $n$ must grow as a tower of tower-height $r-2$. This resolves a missing case in the (as yet unresolved) Erdős–Hajnal conjecture.

The problems in Chapter 3 have a similar but subtly different flavor. Here we consider *random* objects rather than arbitrary or worst-case ones. An example question is: given two random graphs, what is the largest sub-structure (i.e., sub-graph) that they both have in common?

For this and other questions, we are able to give very precise *two-point concentration* results: i.e., show that the answer is, with high probability, one of two adjacent integers.

# Chapter 1

# On multicolor Ramsey numbers of triple system paths of length 3

Let $\mathscr{H}$ be a 3-uniform hypergraph. The multicolor Ramsey number $r_k(\mathscr{H})$ is the smallest integer $n$ such that every coloring of $\binom{[n]}{3}$ with $k$ colors has a monochromatic copy of $\mathscr{H}$. Let $\mathscr{L}$ be the loose 3-uniform path with 3 edges and $\mathscr{M}$ denote the messy 3-uniform path with 3 edges; that is, let $\mathscr{L} = \{abc, cde, efg\}$ and $\mathscr{M} = \{abc, bcd, def\}$. In this chapter we prove $r_k(\mathscr{L}) < 1.54k$ and $r_k(\mathscr{M}) < 1.6k$ for $k$ sufficiently large.

## 1.1 Introduction

Let $r \geq 2$ and consider an $r$-uniform hypergraph $\mathscr{H}$. The multicolor Ramsey number $r_k(\mathscr{H})$ is the minimum $n$ such that every $k$-coloring of $\binom{[n]}{r}$ contains a monochromatic copy of $\mathscr{H}$. The problem of determining the asymptotics of $r_k(\mathscr{H})$ is wide open even for some simple $\mathscr{H}$. Consider, for example, the graph triangle $K_3$. It is known that $r_k(K_3)$ is at least exponential in $k$ and that the limit as $k$ tends to infinity of $r_k(K_3)^{1/k}$ exists. However, the value of this limit remains an open problem; indeed, it is an old \$250 problem of Erdős to determine this limit and a \$100 problem to just determine whether or not this is finite [CG98].

In this work we consider $r_k(\mathscr{H})$ where $\mathscr{H}$ is a 3-edge, 3-uniform path. There are three such hypergraphs: The tight path $\mathscr{T} = \{abc, bcd, cde\}$, the loose path $\mathscr{L} = \{abc, cde, efg\}$, and the messy path $\mathscr{M} = \{abc, bcd, def\}$.

**(a)** Tight Path **(b)** Loose Path **(c)** Messy Path

The tight path was studied in [AGLM14], where it is shown that $2k(1-o(1)) \leq r_k(\mathscr{T}) \leq 2k+3$. The tight path is somewhat different from $\mathscr{L}$ and $\mathscr{M}$ as the tight path has a transversal vertex, i.e., a vertex contained in every edge. Thus, the problem of determining $r_k(\mathscr{T})$ is related to the problem of determining the multicolor Ramsey number of the graph path with 4 vertices, $P_4$. It is known that $2k \leq r_k(P_4) \leq 2k+2$; see [Irv74]. For further results on the multicolor Ramsey numbers of longer graph paths, see [Sár16, DJR17, KS19].

The best known lower bounds on $r_k(\mathscr{L})$ and $r_k(\mathscr{M})$ are

$$r_k(\mathscr{L}) \geq k+6 \quad \text{and} \quad r_k(\mathscr{M}) \geq k+5.$$

The constructions that provide these lower bounds have a common structure. Let $n$ be one less than the bound we are establishing (so $n = k+5$ for the loose path and $n = k+4$ for the messy path). We begin by ordering the vertex set $V = \{v_1, v_2, \ldots, v_n\}$. A triple $v_x v_y v_z$ with $x < y < z$ and $x < k$ is assigned color $x$. The remaining triples are assigned color $k$. The first $k-1$ colors give stars and therefore do not contain copies of either path. The final color is assigned to a complete subhypergraph, but the number of vertices is one fewer than the number of vertices in the path in question. It is believed that these lower bounds give the actual multicolor Ramsey numbers for these hypergraphs [Pol17, ŁP17, ŁP18].

In this work, we provide improvements on the upper bounds on these multicolor Ramsey numbers. The previous best known result for the loose path was $r_k(\mathscr{L}) < 1.975k + 7\sqrt{k}$, which was established by Łuczak and Polcyn [ŁP18]. We are not aware of any discussion of the multicolor Ramsey number of the messy path in literature. Our main results are as follows:

**Theorem 1.1.1.** *If $k$ is sufficiently large then*

$$r_k(\mathscr{L}) < 1.531k.$$

**Theorem 1.1.2.** *If $\varepsilon > 0$ and $k$ is sufficiently large then*

$$r_k(\mathscr{M}) < \left(\frac{10+\sqrt{19}}{9}+\varepsilon\right)k < (1.596+\varepsilon)k.$$

The proofs of these Theorems are similar, and each has two parts. Let $\mathscr{H}$ be $\mathscr{L}$ or $\mathscr{M}$. The first part of the proof is a structural characterization of $\mathscr{H}$-free hypergraphs; in particular, we show that an appropriately chosen core of an $\mathscr{H}$-free hypergraph has a well-organized structure. Such a characterization was first provided by Łuczak and Polcyn for the loose path. See Lemma 19 in [ŁP19]. In the second part of the proof we consider a $k$-coloring of $\binom{[n]}{3}$ which does not contain a monochromatic copy of $\mathscr{H}$. Based on the structural characterization, we introduce a digraph on vertex set $[n]$ in each color. We then proceed to analyze the structure of this colored collection of digraphs to produce the bound on the Ramsey number.

In the case of the loose path, this structural analysis invokes both the Caccetta-Häggkvist Conjecture and the Triangle Removal Lemma for digraphs.

**Conjecture** (Caccetta-Häggkvist [CH78])**.** *If $D$ is a digraph on $n$ vertices with no parallel arcs and minimum in-degree at least $r$ then $D$ has a directed cycle with length at most $\lceil n/r \rceil$.*

We make use of the $r = n/3$ case of this Conjecture. As even this special case is still open, we introduce the following definition.

**Definition.** We say that a constant $\alpha \in [1/3, 2/5)$ is directed triangle sufficient if every oriented graph $D$ with minimum in-degree at least $\alpha|V(D)|$ has a directed cycle of length 3.

Of course, the statement that $1/3$ is directed triangle sufficient is a special case of the Caccetta-Häggkvist Conjecture. The best known result for this special case is that 0.3465 is directed triangle sufficient [HKN17]. Thus, Theorem 1.1.1 is a Corollary of the following Theorem.

**Theorem 1.1.3.** *If $\alpha$ is directed triangle sufficient, $\varepsilon > 0$ and $k$ is sufficiently large then*

$$r_k(\mathscr{L}) < \frac{1+\varepsilon}{1-\alpha}k.$$

Note that if the special case of the Caccetta-Häggkvist holds then $1/3$ is directed triangle sufficient and we would achieve the bound $r_k(\mathscr{L}) < (3/2+\varepsilon)n$.

The organization of this paper is as follows. In the rest of this section we introduce definitions and notation for hypergraphs and graphs. In Section 2, we study the loose path, providing a self-contained proof for a characterization of loose path-free hypergraphs which we then use to prove Theorem 1.1.3. In Section 3, we study the messy path, establishing a characterization of messy path-free hypergraphs, and proving Theorem 1.1.2. Section 4 gives the exact extremal number for the messy path, a result that may be of independent interest.

### 1.1.1 Definitions and notation

We adopt the convention of identifying a hypergraph $\mathscr{H}$ with the edge set of $\mathscr{H}$. We let $V = V(\mathscr{H})$ denote the vertex set of a hypergraph $\mathscr{H}$. A hypergraph $\mathscr{H}$ on vertex set $V$ is $r$-uniform if $\mathscr{H} \subseteq \binom{V}{r}$ where $\binom{V}{r}$ denotes all subsets of $V$ of size $r$. For convenience, we may denote an edge $\{v_1,\ldots,v_r\} \in \mathscr{H}$ by $v_1v_2\ldots v_r$. All hypergraphs considered in this work are 3-uniform.

The multicolor Ramsey number for a hypergraph is closely linked to its extremal number. We define $ex^{(r)}(n,\mathscr{H})$, the *extremal number* of $\mathscr{H}$, to be the maximum number of edges in any $\mathscr{H}$-free $r$-uniform hypergraph on $n$ vertices. Analogously, we define $Ex^{(r)}(n,\mathscr{H})$ to be the *extremal family* of $\mathscr{H}$. This is the set of $\mathscr{H}$-free $r$-uniform hypergraphs on $n$ vertices and $ex^{(r)}(n,\mathscr{H})$ edges. In this work we consider only the case $r = 3$, and so we will simply write $ex(n,\mathscr{H})$ and $Ex(n,\mathscr{H})$ respectively.

Our analysis uses the following concepts. We define the *trace* (sometimes known as a link) of some vertex or set of vertices as

$$Tr(x_1,\ldots,x_k) := \{e \setminus \{x_1,\ldots,x_k\} \mid e \in \mathscr{H}, \{x_1,\ldots,x_k\} \subseteq e\}.$$

The degree of a vertex or set of vertices is then simply $\deg(x_1,\ldots,x_k) := |Tr(x_1,\ldots,x_k)|$. For a 3-uniform hypergraph, we will often refer to $\deg(x,y)$ as the codegree of the pair $x,y$. We define the *m-core* of a hypergraph to be the subhypergraph formed by iteratively removing vertices of degree less than $m$ until every vertex has degree at least $m$ (or the hypergraph is empty).

We also define notation for subhypergraphs. For a hypergraph $\mathscr{H}$ and $U \subseteq V(\mathscr{H})$, we will denote the subhypergraph induced by $U$ by $\mathscr{H}[U] := \{e \in \mathscr{H} : e \subseteq U\}$. We extend the definition and notation for induced subhypergraphs to graphs and digraphs in the natural way.

If $G$ is a graph then the *matching number* of $G$, denoted $\nu(G)$, is the maximum number of edges in a matching in $G$. Furthermore, the *vertex cover number* of $G$, denoted $\tau(G)$, is the minimum number of vertices in a vertex cover of $G$, i.e., a set of vertices which intersects every edge.

## 1.2 Loose Path

The loose path has been studied extensively: a series of papers examines the extremal number, variations of the extremal number, small cases of the multicolor Ramsey number, and asymptotics of the multicolor Ramsey number. The unique extremal $\mathscr{L}$-free hypergraph is the complete star when $n$ is at least 8 [JPR16]. This implies $k+6 \leq r_k(\mathscr{L}) \leq 3k$ for $k \geq 3$ [Jac15, JPR16]. It is known that $r_k(\mathscr{L}) = k+6$ for $k \leq 10$; see [GR12, Jac15, JPR17, PR17b, Pol17].

The previous best known asymptotic upper bound [ŁP18] was $r_k(\mathscr{L}) \leq \lambda k + 7\sqrt{k}$ where $\lambda \approx 1.975$ is a solution to the equation $(\gamma^3 - 3\gamma^2 + 6\gamma - 6)^2 - 72\gamma(2-\gamma)(\gamma-1)^2 = 0$. This result was established by first giving a characterization of loose path-free hypergraphs [ŁP19] and then applying properties of this characterization to each color in a loose-path free $k$-coloring of $\binom{[n]}{3}$ in order to find an upper bound on $n$. We use a similar approach, but arrive at a better bound by encoding more information in a digraph for each color class.

The following Section contains a short, self-contained proof of our characterization of loose path-free hypergraphs. We emphasize that this characterization follows from a similar and stronger result due to Łuczak and Polcyn (namely, Lemma 19 in [ŁP19]). We include a short proof of the characterization in the interest of completeness. We then apply this characterization in Section 2.2 to prove Theorem 1.1.3.

### 1.2.1 Loose Path-Free Hypergraph Characterization

**Theorem 1.2.1.** *If $\mathscr{H}$ is a loose path-free hypergraph and $\mathscr{H}'$ is the 22-core of $\mathscr{H}$ then $\mathscr{H}'$ has the following structure. The vertex set $V(\mathscr{H}')$ has a partition into 3 sets $X, Y, Z$, where the set $X$ is partitioned into sets of size 2 and the set $Z$ is partitioned into sets $(A_v : v \in Y)$. All triples $e$ of the hypergraph $\mathscr{H}'$ have one of the following two forms:*

- *$e \cap X$ is one of the pairs in the partition of $X$ and $|e \cap Y| = 1$*
- *$e \cap Y = \{y\}$ and $e \setminus \{y\} \subseteq A_y$*

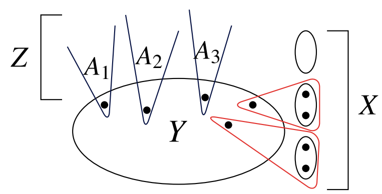



**Figure 1.2.** $\mathscr{L}$-free hypergraph

When we apply this Theorem below we make use of the following definitions.

**Definition 1.2.2.** Each pair in the partition of $X$ is called a locked pair.

**Definition 1.2.3.** Let $\mathscr{H}$ be an $\mathscr{L}$-free hypergraph and let $\mathscr{H}'$ be the 22-core of $\mathscr{H}$. We call the triples in $\mathscr{H} \setminus \mathscr{H}'$ stray triples (or removed triples).

The remainder of this subsection is a proof of Theorem 1.2.1. Throughout the proof we let $\mathscr{H}$ be a loose path-free hypergraph and $\mathscr{H}'$ be the 22-core of $\mathscr{H}$.

**Lemma 1.2.4.** *The matching number $\nu(Tr_{\mathscr{H}'}(v)) \neq 2, 3$ for all vertices $v \in V(\mathscr{H}')$.*

*Proof.* Let $v$ be a fixed vertex and assume for the sake of contradiction that the matching number of the trace of $v$ in $\mathscr{H}'$ is 2 or 3. Let $M$ be a maximal matching in $Tr_{\mathscr{H}'}(v)$ and let $\mathscr{M}$ be the union of the edges in $M$. As the vertex set $\mathscr{M}$ contains at most 15 edges, there is an edge $e$ of $Tr_{\mathscr{H}'}(v)$ that is not contained in $\mathscr{M}$. As $M$ is a maximal matching, $e$ intersects $\mathscr{M}$ in one vertex. Let $u$ be the vertex in $e$ that is not in $\mathscr{M}$. As any edge of $Tr_{\mathscr{H}'}(v)$ that contains $u$ must also intersect $\mathscr{M}$, there are at most 6 such edges (this count includes $e$ itself). As $u$ is in at least $22 > \binom{6}{2} + 6$ edges in $\mathscr{H}'$, there is a triple $uyz \in \mathscr{H}'$ such that at least one of $y$ and $z$ is not in $\mathscr{M}$ and neither $y$ nor $z$ is $v$. Consider such a triple $uyz \in \mathscr{H}'$. A loose path appears among two edges of $M \cup \{e\}$ (expanded to triples by including $v$) and $uyz$. □

We are now ready to identify the first part of the structure defined in Theorem 1.2.1. We define $Y$ to be the set of vertices $y$ such that the matching number of $Tr_{\mathscr{H}'}(y)$ is at least 4.

**Lemma 1.2.5.** *If $y \in Y$ and a triple $abc \in \mathscr{H}'$ intersects an edge $e$ of $Tr_{\mathscr{H}'}(y)$ then either $e \subset abc$ or $y \in abc$.*

*Proof.* Suppose $y \notin abc$ and $|e \cap abc| = 1$. Let $M$ be a maximum matching of $Tr_{\mathscr{H}'}(y)$. As this matching has at least 4 edges, the edge set $M \cup \{e\}$ contains an edge that intersects $abc$ in 1 vertex and an edge that does not intersect $abc$ and is disjoint from the first edge. These two edges (expanded to triples by including $y$) and $abc$ form a loose path. □

We now note that Lemma 1.2.5 implies that no triple in $\mathscr{H}'$ contains more than one vertex of $Y$. Indeed, assume for the sake of contradiction that $y_1y_2a \in \mathscr{H}'$ where $y_1, y_2 \in Y$. Consider a triple that contains $y_1$. Such a triple intersects the edge $y_1a$ of $Tr_{\mathscr{H}'}(y_2)$. It follows that either the triple contains $y_2$ or the triple contains $a$. This then implies that the matching number of $Tr_{\mathscr{H}'}(y_1)$ is at most 2, which is a contradiction.

**Lemma 1.2.6.** *Let $y \in Y$. Every connected component of $Tr_{\mathscr{H}'}(y)$ has either at most 2 vertices or at least 23 vertices. Furthermore, if $C$ is the vertex set of a component with at least 23 vertices and $abc \in \mathscr{H}'$ intersects $C$ then $y \in abc$.*

*Proof.* Consider a connected component with at least 3 vertices. Let $u$ be a vertex of this component of maximum degree; note that $\deg(u) \geq 2$. The triples of $\mathscr{H}'$ that contain $u$ either contain $y$ and therefore correspond to edges of $Tr_{\mathscr{H}'}(y)$ or contain all of the neighbors of $u$ in $Tr_{\mathscr{H}'}(y)$ (by Lemma 1.2.5). The latter condition cannot be satisfied if the degree of $u$ in $Tr_{\mathscr{H}'}(y)$

is greater than 2. On the other hand, if the degree of $u$ in $Tr_{\mathscr{H}'}(y)$ is 2 then we have at most 3 triples of $\mathscr{H}'$ that contain $u$, which is a contradiction. We conclude that the degree of $u$ in $Tr_{\mathscr{H}'}(y)$ is its degree in $\mathscr{H}'$, which is at least 22, and so there are at least 23 vertices in the component.

To prove the second assertion in the Lemma, assume for the sake of contradiction that the triple $abc \in \mathscr{H}'$ contains a vertex $u$ of a component $C$ that has at least 23 vertices but does not contain $y$. Then, by repeated application of Lemma 1.2.5, we see that $abc$ contains all vertices of $C$. As $3 < 23$ this is a contradiction. □

We say that for a vertex $y \in Y$, a connected component of $Tr_{\mathscr{H}'}(y)$ is large if it has at least 23 vertices.

We are now ready to identify the other parts of the vertex partition set forth in Theorem 1.2.1. For each vertex $y \in Y$, let $A_y$ be the union of the vertex sets of the large components of $Tr_{\mathscr{H}'}(y)$. Set $Z := \bigsqcup_{y \in Y} A_y$. As no triple in $\mathscr{H}'$ contains more than one vertex of $Y$, the sets $Y$ and $Z$ are disjoint. Set $X = V(\mathscr{H}') \setminus (Y \cup Z)$.

We have our partition $X, Y, Z$ and the partition of $Z$ into $(A_v : v \in Y)$. We now look to partition $X$. Let $x \in X$. It follows from Lemma 1.2.4 and the definition of $Y$ that $Tr_{\mathscr{H}'}(x)$ is a star with at least 22 edges. Let $x'$ be the center of this star. Note that $x' \notin Y$: if $x' \in Y$ then $x$ itself would be in a large component of $A_{x'}$ and so $x$ would be in $Z$. Furthermore, $x' \notin Z$ as this would imply – by Lemma 1.2.6 – that all at least 22 triples containing $x$ must contain both $x'$ and some fixed element of $Y$ (of course, there is only one such triple). It follows that $x' \in X$. We conclude that $X$ can be partitioned into a collection of pairs $xx'$ with the property that every triple of $\mathscr{H}'$ that contains one vertex in such a pair also contains the other vertex in the pair. Finally, note that the third vertex in such a triple must be in the set $Y$. Thus, triples intersecting $X$ are as stated.

### 1.2.2 Multicolor Ramsey number for the loose path: Proof of Theorem 1.1.3

In this Section we prove Theorem 1.1.3. The proof invokes the Caccetta-Häggkvist Conjecture and the Removal Lemma for directed triangles. We recall the following for reference.

**Definition 1.2.7.** We say that a constant $\alpha \in [1/3, 2/5)$ is directed triangle sufficient if every oriented graph $D$ with minimum in-degree at least $\alpha|V(D)|$ has a directed cycle of length 3.

**Theorem 1.2.8** (Alon, Shapira [AS04], Digraph Removal)**.** *For every fixed $\delta, h$, there is a positive constant $c(h,\delta)$ with the following property. If $H$ is a fixed digraph on $h$ vertices and $G$ is a digraph on $n$ vertices, where $n$ is sufficiently large, with the property that upon the removal of at most $\delta n^2$ arcs $G$ still contains a copy of $H$ then $G$ contains at least $c(h,\delta)n^h$ copies of $H$.*

Let $\alpha$ be directed triangle sufficient and let $\varepsilon > 0$ be a small constant. Suppose

$$n = \frac{1+\varepsilon}{1-\alpha}k,$$

and assume for the sake of contradiction that there is a loose path-free $k$-coloring of $\binom{[n]}{3}$. Fix such a coloring and let $C$ be the set of colors. For each color $c \in C$, let $\mathscr{H}_c$ be the collection of triples colored with color $c$. We apply Theorem 1.2.1. For each $\mathscr{H}_c$, we let $\mathscr{H}_c'$ be the 22-core of $\mathscr{H}_c$ and we let $X_c, Y_c, Z_c$ be the partition of the vertex set given by Theorem 1.2.1. Furthermore, for each vertex $v$ in the set $Y_c$ let $A_{v,c}$ be the set $A_v$ given by Theorem 1.2.1. We define a colored multidigraph $M$ on the vertex set $V = [n]$ as follows. The directed arc $(u,v)$ appears in the multidigraph with color $c$ if $u \in A_{v,c}$. For a specific color, we denote this arc by $(u,v,c)$. We will also include both $(u,v,c), (v,u,c)$ in the multidigraph if $u,v$ is a locked pair in color $c$ (recall Definition 1.2.2). Our main focus in the proof will be on the pairs of vertices that have arcs of $M$ going in only one direction; this will be most pairs.

We define the in and out-degrees of the colored multidigraph $M$ as follows:

$$m_c^-(v) = |\{u \in [n] : (u,v,c) \in M\}|, \quad m^-(v) = \sum_{c \in C} m_c^-(v)$$

$$m_c^+(v) = |\{u \in [n] : (v,u,c) \in M\}|, \quad m^+(v) = \sum_{c \in C} m_c^+(v)$$

Note that $m^+(v) \leq k$ as $v$ can appear either in $A_{y,c}$ for at most one vertex $y$ or in a single locked pair in color $c$.

**Lemma 1.2.9.** *At most $O(k)$ pairs of vertices $\{u,v\}$ have the property that neither $(u,v)$ nor $(v,u)$ appears as an arc in any color in $M$.*

*Proof.* Note that if neither $(u,v,c)$ nor $(v,u,c)$ appears then at most one triple containing $\{u,v\}$ appears in $\mathscr{H}_c'$. So, if neither $(u,v)$ nor $(u,v)$ appears in any color then at least $n-k-2$ triples containing $u$ and $v$ are stray triples. As there are at most $21nk$ stray triples across the $k$ colors, we see that the number of pairs $\{u,v\}$ that span no arc of $M$ is at most $\frac{21nk}{n-k-2} = O(k)$. □

We will refer to pairs $\{u,v\}$ such that neither $(u,v)$ nor $(v,u)$ appears in $M$ as *uncovered pairs*.

We define an oriented graph $D$ on $[n]$ as follows

- $(v,u) \in D$ if $(v,u) \in M$ and $(u,v) \notin M$
- $d^+(v) = |\{u \in [n] : (v,u) \in D\}|$
- $d^-(v) = |\{u \in [n] : (u,v) \in D\}|$

Note that almost all vertices $v$ have $d^-(v) > n-k-o(k)$: the out-degrees in $M$ are at most $k$ and there are $O(k)$ uncovered pairs while the remaining arcs at a vertex are then in-arcs in $D$.

We now apply Directed Triangle Removal. Set

$$\delta = \varepsilon^2/17$$

and consider any digraph $D'$ formed by deleting $\delta n^2$ arcs of $D$. We claim that $D'$ has a directed triangle. Let $X$ be the set of vertices $x$ such that the number of arcs directed into $x$ that are deleted plus the number of uncovered pairs of $M$ incident with $x$ is at least $\varepsilon n/4$. We claim that $|X| < \varepsilon n/4$. Indeed, if this bound does not hold then the number of deleted arcs plus twice the number of uncovered pairs of $M$ is at least $(\varepsilon^2/16)n^2$, which cannot be the case for $k$ sufficiently large by Lemma 1.2.9. Now consider the induced digraph $D'[[n] \setminus X]$. In-degrees in this digraph are at least $n-k-1-\varepsilon n/4-\varepsilon n/4$ as within $[n]\setminus X$, we have max outdegree $k$ in $M$ and at most $\varepsilon n/4$ uncovered pairs and deleted in-arcs at a vertex. Thus the minimum in-degree in this induced sub-digraph is at least

$$n-k-\frac{n\varepsilon}{2}-1=n\left(1-\frac{1-\alpha}{1+\varepsilon}-\frac{\varepsilon}{2}\right)-1=n\left(\frac{\alpha+\varepsilon/2-\varepsilon^2/2}{1+\varepsilon}\right)-1>\alpha n,$$

where we use $\alpha<2/5$ in the last inequality. Now, since $\alpha$ is directed triangle sufficient, we conclude that $D'[[n]\setminus X]$ contains a directed triangle.

Directed Triangle Removal implies that $D$ contains $\Omega(n^3)$ directed triangles. Consider such a triangle $xyz$ and the color $c$ such that $xyz\in\mathscr{H}_c$. No pair among $xyz$ can be a locked pair for this color as we have arcs in only one direction. Furthermore the triple $xyz$ cannot be contained in one of the stars in $\mathscr{H}_c'$ as this would require a vertex of in-degree two in the color $c$ digraph induced on $xyz$. We conclude that $xyz$ is a stray triple in color $c$. But there are at most $21nk$ stray triples, and if $k$ is sufficiently large, we can find some directed triangle which is not covered by a stray triple and thus is uncolored in our coloring. This is a contradiction.

*Remark* 1.2.10. Applying that $\alpha=.3465$ is directed triangle sufficient [HKN17], we have that $r_k(\mathscr{L})\leq 1.531k$ for $k$ sufficiently large.

## 1.3 Messy Path

The messy path is the hypergraph $\mathscr{M}=\{abc,bcd,def\}$. Extremal results for collections of hypergraphs containing $\mathscr{M}$ are studied in [FJK$^+$21], where the messy path is $P_3(1,2)$ or $P_3(2,1)$, and in [FO11], where the messy path is a $(2,1)$-cluster. In [FO11], it is shown that for sufficiently large $n$, $ex(n,\mathscr{M})=\binom{n-1}{2}$ with the unique extremal hypergraph being a complete star. In Section 4 we find the extremal number for all $n$. This bound on the extremal number implies $r_k(\mathscr{M})\leq 3k$ if $k\geq 3$.

This Section is dedicated to the proof of Theorem 1.1.2. The outline of the proof is the same as the for the loose path. We begin with a structural characterization of $\mathscr{M}$-free hypergraphs. We then use this characterization to define a colored multidigraph associated with a $\mathscr{M}$-free

$k$-coloring of $\binom{[n]}{3}$ and proceed to establish our upper bound on $r_k(\mathscr{M})$.

### 1.3.1 Messy Path-Free Hypergraph Characterization

We begin with our characterization of messy path-free hypergraphs. Note that, like our characterization of $\mathscr{L}$-free hypergraphs, this characterization features disjoint stars. However, the rest of the characterization is less well-behaved and hence more challenging in the application that follows.

**Theorem 1.3.1.** *Let $\mathscr{H}$ be a messy path-free hypergraph, and let $\mathscr{H}'$ be the 13-core of $\mathscr{H}$. The vertex set $V(\mathscr{H}')$ has a partition into 3 sets $X, Y, Z$ and the set $Z$ has a partition into sets $(A_v : v \in Y)$ such that all triples $e$ of the hypergraph $\mathscr{H}'$ have one of the following two forms:*

- *$e \subseteq X \cup Y$ where $\mathscr{H}[X \cup Y]$ is a partial Steiner Triple System*
- *$e \cap Y = \{y\}$ and $e \setminus \{y\} \subseteq A_y$*

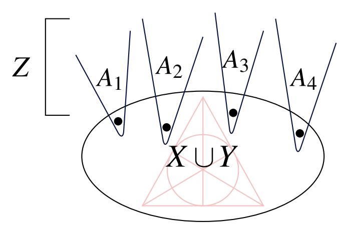



**Figure 1.3.** $\mathscr{M}$-free hypergraph

In the application of Theorem 1.3.1 we make use of the following definition.

**Definition 1.3.2.** If $\mathscr{H}$ is an $\mathscr{M}$-free hypergraph and $\mathscr{H}'$ is the 13-core of $\mathscr{H}$ then the triples in $\mathscr{H} \setminus \mathscr{H}'$ will be called stray triples (or removed triples).

The remainder of this subsection is a proof of Theorem 1.3.1. Throughout the proof, we let $\mathscr{H}$ be an $\mathscr{M}$-free hypergraph and $\mathscr{H}'$ be the 13-core of $\mathscr{H}$. Recall that $F(a,2)$ refers to the 3-uniform hypergraph with vertex set $\{x_1, \ldots, x_a, y_1, y_2\}$ and edge set $\{x_i y_1 y_2 : i \in [a]\}$. We refer to the vertices $x_1, \ldots, x_a$ as petals of the $F(a,2)$ and $y_1, y_2$ as the center.

**Lemma 1.3.3.** *The vertex cover number $\tau(Tr_{\mathscr{H}'}(v)) \neq 2, 3$ for all vertices $v \in V(\mathscr{H}')$.*

*Proof.* Let $v$ be a fixed vertex. Assume for the sake of contradiction that the vertex cover number of the trace of $v$ in $\mathscr{H}'$ is 2 or 3. Let $U$ be a minimal vertex cover in $Tr_{\mathscr{H}'}(v)$. Since $Tr_{\mathscr{H}'}(v)$ has at least 13 edges, one of these vertices has degree at least 5, say $u$. Let $x \in U$ with $x \neq u$. Note that by definition, there exists an edge $xy$ where $y \neq u$. Then $u$ has a neighbor $a$ in $Tr_{\mathscr{H}'}(v)$ which is not in $U$ and is neither $x$ nor $y$. Consider a triple $afg$.

- If $u, v \notin afg$ then $\mathscr{M}$ appears. (Consider $afg$ together with the triple corresponding to a path in $Tr_{\mathscr{H}'}(v)$ centered at $u$.)
- If $v \notin afg$, $u \in afg$, and $\{x,y\} \cap \{f,g\} = \emptyset$ then $\mathscr{M}$ appears. (Consider $afg = aug, auv, vxy$.)

So we may assume that such triples do not appear in $\mathscr{H}'$. Now we observe that there are at most 3 triples that contain $a$ and $v$ by the bound on the vertex cover number. There are at most two triples that contain $a$ and $u$ as such triples contain $x$ or $y$. Thus, $a$ is in at most 5 triples in $\mathscr{H}'$, which is a contradiction. □

We can now start to identify the parts in the vertex partition set forth in Theorem 1.3.1. Let $Z$ be the set of vertices whose traces in $\mathscr{H}'$ have vertex cover number 1.

**Lemma 1.3.4.** *If $z \in Z$ and $y$ is the center of the star in $Tr_{\mathscr{H}'}(z)$, then $\tau(Tr_{\mathscr{H}'}(y)) \geq 4$. Furthermore, if $u$ is any leaf of the star in $Tr_{\mathscr{H}'}(z)$, then $\tau(Tr_{\mathscr{H}'}(u)) = 1$.*

*Proof.* Assume for the sake of contradiction that $\tau(Tr_{\mathscr{H}'}(y)) = 1$. Then, $y, z$ form the center of a $F(13,2)$ in $\mathscr{H}'$. Note that any petal of this $F(13,2)$ has degree 1 or else there is a messy path, a contradiction.

Assume for the sake of contradiction that some leaf $u$ has $\tau(Tr_{\mathscr{H}'}(u)) \geq 4$. Then, we can find a triple in $\mathscr{H}'$ that contains $u$ and neither $y$ nor $z$. This triple together with two triples corresponding to a path of two edges in $Tr_{\mathscr{H}'}(z)$ forms a messy path, a contradiction. □

We now let $Y$ be the set of vertices $y$ for which there is $z \in Z$ such that $y$ is the center of the star in $Tr_{\mathscr{H}'}(z)$. For $y \in Y$, let $A_y$ be the set of $z \in Z$ for which $y$ is the center of the star in $Tr_{\mathscr{H}'}(z)$. Finally, let $X = V(\mathscr{H}') \setminus (Y \cup Z)$. Note that in $X, Y$, every vertex has a trace with vertex cover number at least 4. Further note that $(A_y : y \in Y)$ partitions $Z$.

**Lemma 1.3.5.** *The hypergraph $\mathscr{H}'[X \cup Y]$ is a partial Steiner Triple System.*

*Proof.* Assume for the sake of contradiction that there are triples $wxy, xyz \in \mathscr{H}'[X \cup Y]$. Since $Tr_{\mathscr{H}'}(w)$ has vertex cover number at least 4, there exists an edge in $Tr_{\mathscr{H}'}(w)$ not containing $x, y, z$. The triple corresponding to this edge together with $wxy, xyz$ form a messy path. □

Theorem 1.3.1 follows from Lemmas 1.3.4 and 1.3.5.

### 1.3.2 Multicolor Ramsey number for the Messy Path: Proof of Theorem 1.1.2

The proof of Theorem 1.1.2 uses a digraph structure analogous to the structure introduced in the proof of Theorem 1.1.3. In that proof it was sufficient to find a cubic number of directed triangles. This is not sufficient for the messy path as a cubic number of triples can be contained in the 'Steiner parts' of the color classes. So we make a more subtle argument that takes more of the structure into account.

Throughout this section we consider a messy path-free $k$-coloring of $\binom{[n]}{3}$ where $k$ is assumed to be large. We establish upper bounds on $n$ that imply Theorem 1.1.2. Let $C$ be the set of colors. For each color $c \in C$, let $\mathscr{H}_c$ be the collection of triples colored with color $c$. We apply Theorem 1.3.1. For each $\mathscr{H}_c$, we let $\mathscr{H}_c'$ be the 13-core of $\mathscr{H}_c$ and we let $X_c, Y_c, Z_c$ be the partition of the vertex set given by Theorem 1.3.1. Furthermore, for each vertex $v$ in the set $Y_c$ let $A_{v,c}$ be the set $A_v$ given by Theorem 1.3.1. We define a colored multidigraph $M$ on the vertex set $V = [n]$ as follows. The directed arc $(u, v)$ appears in the multidigraph with color $c$ if $u \in A_{v,c}$, i.e., $u$ points to $v$ if $u$ is contained in the body of a star centered at $v$. For a specific color, we denote this arc by $(u, v, c)$.

Define

$$m_c^-(v) = |\{u \in [n] : (u, v, c) \in M\}|, \quad m^-(v) = \sum_{c \in C} m_c^-(v)$$

$$m_c^+(v) = |\{u \in [n] : (v, u, c) \in M\}|, \quad m^+(v) = \sum_{c \in C} m_c^+(v)$$

Note that $m^+(v) \le k$ as $v$ can be in the body of at most one star in each $\mathscr{H}'_c$.

**Lemma 1.3.6.** *There are at most $O(k)$ pairs of vertices $\{u,v\}$ such that neither $(u,v)$ nor $(v,u)$ appear as an arc in any color in $M$.*

*Proof.* Note that if neither $(u,v,c)$ nor $(v,u,c)$ appears then there is at most one triple in $\mathscr{H}'_c$ that contains the pair $\{u,v\}$. Therefore $k$ plus the number of stray triples containing $\{u,v\}$ is at least $n-2$, and such a pair $\{u,v\}$ is contained in at least $n-k-2$ stray triples. As there are at most $12nk$ stray triples across the colors, there are at most $\frac{12nk}{n-k-2} = O(k)$ such pairs of vertices. □

We need further notation to extract more refined information from the colored multidigraph $M$. For each vertex $v$ let $s(v)$ be the number of colors $c$ for which $v \in X_c \cup Y_c$; in other words, $s(v)$ is the number of colors $c$ for which $v$ is in the Steiner triple system for color $c$. Define

$$S = \sum_{v\in[n]} s(v).$$

Note that $v \in A_{u,c}$ implies that $v$ is not part of a partial Steiner triple system in $\mathscr{H}'_c$. Hence, we have

$$m^+(v) + s(v) \le k \quad \text{and} \quad |M| = \sum_{v\in[n]} m^+(v) \le \sum_{v\in[n]} k - s(v) = nk - S. \tag{1.1}$$

We will also want to keep track of stray triples. For $u,v \in [n]$ define

$$\xi_{uv} = \left|\left\{z : \exists c \in C \text{ such that } uvz \in \mathscr{H}_c \setminus \mathscr{H}'_c\right\}\right|$$
$$\xi_{u} = \left|\left\{\{y,z\} : \exists c \in C \text{ such that } uyz \in \mathscr{H}_c \setminus \mathscr{H}'_c\right\}\right|.$$

Note that $\frac{1}{3}\sum_{u\in[n]} \xi_u, \frac{1}{3}\sum_{\{u,v\}\in\binom{[n]}{2}} \xi_{uv} \le 12nk$. Next, we categorize pairs of vertices.

- We say that a vertex pair $\{u,v\}$ is a two-cycle pair if $(u,v),(v,u) \in M$. Let $t(u)$ be the number of two-cycle pairs that contain $u$. We also define

$$T = \frac{1}{2}\sum_{u\in[n]} t(u) = \text{ number of two-cycle pairs of vertices,} \qquad \bar{t} = \frac{2T}{n}$$

- We say a pair $\{u,v\}$ is a parallel pair if $\{u,v\}$ is not a two-cycle pair and there are at least two arcs in $M$ contained in the pair, i.e., $(u,v,c)$ and $(u,v,c')$ appear or $(v,u,c)$ and $(v,u,c')$

appear for $c, c'$ distinct. Let $p^+(u)$ and $p^-(u)$ be the number of parallel pairs that are directed out of and into $u$, respectively. We also let:

$$P = \sum_{u \in [n]} p^+(u) = \text{ number of parallel pairs of vertices,} \qquad \hat{p} = \frac{P}{n-k}$$

- We say that a pair $\{u, v\}$ is a solo pair if either $(u, v, c)$ or $(v, u, c)$ for some color $c$ is the only arc on the pair. Let $q^+(u)$ and $q^-(u)$ be the number of solo arcs that are directed out of and into $u$ respectively.
- We say that a pair $\{u, v\}$ is an uncovered pair if neither $(u, v)$ nor $(v, u)$ appears in $M$.

Now, Lemma 1.3.6 implies that at least $\binom{n}{2} - O(k)$ pairs of vertices are covered with at least one arc, and each two-cycle pairs and each parallel pair requires at least one additional arc. It follows that we have

$$P + T \leq |M| - \binom{n}{2} + O(k) \leq nk - S - \binom{n}{2} + O(k). \tag{1.2}$$

Below we establish two upper bounds on $n$, one which is a function of $P$ and another that is a function of $T$. These bounds taken together with (1.2) imply Theorem 1.1.2.

We need additional notation. We define an oriented graph $D$ on $[n]$, which defines the one-way out-neighborhood and in-neighborhood $D^+(v), D^-(v)$ (and out-degree and in-degree) for a vertex $v \in [n]$ as follows

- $(v, u) \in D$ if $(v, u) \in M$ and $(u, v) \notin M$
- $D^+(v) = \{u \in [n] : (v, u) \in D\}, d^+(v) = |D^+(v)|$
- $D^-(v) = \{u \in [n] : (u, v) \in D\}, d^-(v) = |D^-(v)|$

Note that $D$ consists of arcs corresponding to solo pairs and parallel pairs.

Consider a fixed vertex $v$. Observe that for any other vertex $u$ we either have $(v, u) \in M$ (and $u$ is counted by $m^+(v)$), $(u, v) \in M$ and $(v, u) \notin M$ (and $u$ is counted by $d^-(v)$), or $\{u, v\}$ is uncovered (and there are at least $n-k$ stray triples that contain $uv$). Therefore $m^+(v) + d^-(v) + \xi_v/(n-k) \geq n-1$. Applying (1.1) it follows that we have

$$k - s(v) + p^-(v) + q^-(v) \geq m^+(v) + d^-(v) \geq n - \xi_v/(n-k) - 1 \tag{1.3}$$

With these preliminary observations in hand, we are now ready to state two key Lemmas. The first Lemma gives a bound on $n$ in terms of $\bar{t}$ while the second Lemma gives a bound on $n$ in terms of $\hat{p}$. We then complete the proof by combining these Lemmas and the bound on $P+T$ given in (1.2).

**Lemma 1.3.7.** *If $\varepsilon > 0$ and $k$ is sufficiently large then*

$$n \leq \frac{4k}{3} + \bar{t} + \varepsilon k$$

*Proof.* First note that

$$|D| \geq \binom{n}{2} - T - O(k)$$

as every pair of vertices is either a solo pair, a parallel pair, a two-cycle pair, or an uncovered pair. Then note that $\sum_{v\in[n]} d^+(v) + \frac{d^-(v)}{2} = \frac{3}{2}|D|$ and so

$$\mathbb{E}_{v\in[n]}\left[d^+(v) + \frac{d^-(v)}{2} - \frac{2\xi_v}{\varepsilon k}\right] \geq \frac{3(n-\bar{t})}{4} - O(1/\varepsilon).$$

Let $v$ be a vertex that maximizes $d^+(v) + \frac{d^-(v)}{2} - \frac{2\xi_v}{\varepsilon k}$.

Note that we may assume $d^-(v) > \varepsilon k$ as otherwise we have

$$k > d^+(v) \geq \frac{3(n-\bar{t})}{4} - O(1/\varepsilon) - \frac{\varepsilon k}{2}$$

and the desired bound follows.

By Lemma 1.3.6, we have $\left|\{\{u,w\} \in \binom{D^-(v)}{2} : (u,w) \text{ or } (w,u) \in M\}\right| \geq \binom{d^-(v)}{2} - O(k)$. Therefore

$$\mathbb{E}_{u\in D^-(v)}\left[\left|\{w \in D^-(v) : (u,w) \in M\}\right| - \xi_{uv}\right] \geq \frac{d^-(v)}{2} - O(1/\varepsilon) - \frac{2\xi_v}{\varepsilon k}.$$

Consider a vertex $u \in D^-(v)$ that maximizes $|\{w \in D^-(v) : (u,w) \in M\}| - \xi_{uv}$. Note that for every vertex $x \in D^+(v)$, the triple $uvx$ appears in some $\mathscr{H}_c$ in one of the following ways: $uvx$ is contained in the partial Steiner triple system on $X_c \cup Y_c$; $u,v \in A_{x,c}$ so that $(u,x,c)$ and $(v,x,c)$ both appear in $M$; or $uvx$ is one of the stray triples. (Note that we are making of the use of the fact that $D$ only includes arcs that are not in 2-cycles in $M$. In particular $(x,v)$ and $(v,u)$ do not appear in $M$.) Therefore, as $D^-(v)$ and $D^+(v)$ are disjoint,

$$s(u)+m^+(u)-\left|\left\{w\in D^-(v):(u,w)\in M\right\}\right|+\xi_{uv}\geq d^+(v).$$

Recalling $k\geq s(u)+m^+(u)$ from (1.1) we have

$$\begin{aligned}k &\geq d^+(v)+\left|\left\{w\in D^-(v):(u,w)\in M\right\}\right|-\xi_{uv}\geq d^+(v)+\frac{d^-(v)}{2}-O(1/\varepsilon)-\frac{2\xi_v}{\varepsilon k}\\ &\geq\frac{3(n-\bar{t})}{4}-O(1/\varepsilon).\end{aligned}$$

Rearranging and letting $k$ be sufficiently large, we have

$$n\leq\frac{4k}{3}+\bar{t}+\varepsilon k$$ □

**Lemma 1.3.8.** *If* $0<\varepsilon<0.01$ *and k is sufficiently large then*

$$n\leq\max\left\{1.59k,\frac{3}{2}k+\frac{\hat{p}}{2}+\varepsilon k\right\}$$

*Proof.* We say that a vertex $v$ is light if $\max_{c\in C}m_c^-(v)\leq n/2$. Note that since each color contributes at most one vertex which is not light, there are at least $n-k$ light vertices. Let $L$ be the collection of light vertices. By (1.3), we have

$$\mathbb{E}_{v\in L}\left[q^-(v)-s(v)-\frac{6\xi_v}{\varepsilon k}\right]\geq n-k-\hat{p}-O(1/\varepsilon).$$

Let $v$ be a vertex of $L$ that maximizes $q^-(v)-s(v)-\frac{6\xi_v}{\varepsilon k}$.

Observe that, appealing to (1.2), we have

$$\hat{p}\leq\frac{nk-n^2/2+O(k)}{n-k}.$$

Assuming that $n\geq 1.59k$ and noting that the above bound is decreasing in $n$ (and consequently maximized when $n=1.59k$), we have

$$q^-(v)-s(v)-\frac{6\xi_v}{\varepsilon k}\geq .59k-\frac{1.59k^2-(1.59k)^2/2}{.59k}-O(1/\varepsilon)\geq .0375k-O(1/\varepsilon)\geq 2\varepsilon k \quad (1.4)$$

and so $q^-(v)-s(v)$ is linear in size.

Let $B$ be the set of vertices $u$ such that $(u,v)$ is a solo arc in $M$. Note that $|B|=q^-(v)$. We now consider cases based on the colors on the solo arcs directed from $B$ into $v$. For each color $c$ let $B_c$ be the set of vertices $x \in B$ such that $xv$ is colored $c$.

**Case 1.** *Some $c$ has $\varepsilon k < |B_c|$.*

Let $w \in B_c$ such that $\xi_{wv} \le 2\xi_v/(\varepsilon k)$. Consider the triples of the form $wvz$ where $z \in B \setminus B_c$. As $w$ and $z$ point to $v$ with solo arcs of different colors, such a triple must be covered by stray triples or a triple in a partial Steiner triple system. Thus,

$$s(v) + 2\xi_v/(\varepsilon k) \ge s(v) + \xi_{wv} \ge q^-(v) - |B_c|.$$

(Note that if $B = B_c$, then the above inequality just states $s(v) + 2\xi_v/(\varepsilon k) \ge 0$.) Therefore,

$$|B_c| \ge q^-(v) - s(v) - 2\xi_v/(\varepsilon k) \ge n - k - \hat{p} + 4\xi_v/(\varepsilon k) - O(1/\varepsilon).$$

Now note Lemma 1.3.6 implies that $\left|\{\{u,z\} \in \binom{B_c}{2} : (u,z) \text{ or } (z,u) \in M\}\right| \ge \binom{|B_c|}{2} - O(k)$ so we have

$$\mathbb{E}_{u \in B_c}\left[|\{z \in B_c : (u,z) \in M\}| - \xi_{uv}\right] \ge \frac{|B_c|}{2} - O(1/\varepsilon) - \frac{2\xi_v}{\varepsilon k}.$$

Consider a vertex $u \in B_c$ that maximizes $|\{z \in B_c : (u,z) \in M\}| - \xi_{uv}$. Let $C$ be the set of vertices $y$ such that $(y,v,c) \notin M$. Note that for every vertex $y \in C$, the triple $uvy$ appears in some $\mathscr{H}_{c'}$ in one of the following ways: $uvy$ is contained in the partial Steiner triple system on $X_{c'} \cup Y_{c'}$; $u, v \in A_{y,c'}$ so that $(u,y,c'), (v,y,c') \in M$; or $uvy$ is one of the stray triples. Therefore,

$$s(u) + m^+(u) - |\{z \in B_c : (u,z) \in M\}| + \xi_{uv} \ge |C|.$$

Since $k \ge s(u) + m^+(u)$ from (1.1), we have

$$\begin{aligned} k &\ge |C| + |\{z \in B_c : (u,z) \in M\}| - \xi_{uv} \ge |C| + \frac{|B_c|}{2} - O(1/\varepsilon) - \frac{2\xi_v}{\varepsilon k} \\ &\ge \frac{n}{2} + \frac{n-k-\hat{p}}{2} - O(1/\varepsilon), \end{aligned}$$

where the bound $|C| \ge n/2$ follows from the fact that $v$ is a light vertex. Rearranging and letting $k$ be sufficiently large, we have

$$n \le \frac{3k}{2} + \frac{\hat{p}}{2} + \varepsilon k,$$

as desired.

**Case 2.** *Every color appears on at most $\varepsilon k$ solo arcs from $B$ to $v$.*

Consider the triples containing $v$ and two vertices in $B$. The solo arcs partition $B$ so that the $\ell$ colors on these arcs color at most $\sum_{i=1}^{\ell}\binom{x_i}{2}$ of these triples, where $x_i$ is the number of vertices with solo arcs in the $i^{\text{th}}$ color. Noting that $x_i \le \varepsilon k$ and $\sum_{i=1}^{\ell} x_i = q^-(v)$, we have that by convexity,

$$\sum_{i=1}^{\ell}\binom{x_i}{2} \le \sum_{i=1}^{\ell}\frac{x_i^2}{2} \le \frac{q^-(v)}{\varepsilon k}\cdot\frac{(\varepsilon k)^2}{2} \le \frac{q^-(v)\varepsilon k}{2}$$

Furthermore, the number of triples in the partial Steiner triple systems that contain $v$ and two vertices from $B$ is at most $s(v)\frac{q^-(v)}{2}$, since each partial Steiner triple system contributes a matching in the trace of $v$. The only other triples containing $v$ and two vertices in $B$ are the $\xi_v$ stray triples containing $v$. Then, we have

$$\binom{q^-(v)}{2} \le \frac{q^-(v)\varepsilon k}{2} + \frac{s(v)q^-(v)}{2} + \xi_v$$

so $q^-(v) - s(v) \le \varepsilon k + 1 + \frac{2\xi_v}{q^-(v)}$ and

$$q^-(v) - s(v) - \frac{6\xi_v}{\varepsilon k} \le \varepsilon k + 1.$$

This contradicts (1.4) when $k$ is sufficiently large. □

We are now ready to complete the proof. Lemmas 1.3.7 and 1.3.8 imply that if either $\hat{p}$ or $\bar{t}$ is sufficiently small then the desired upper bound on $n$ follows. Observe that (1.2) implies

$$\hat{p}(n-k) + \frac{n\bar{t}}{2} \le nk - \frac{n^2}{2} + O(k). \tag{1.5}$$

Given this linear relationship, we see that $n$ is at most the bound determined by setting the expressions in the two Lemmas equal. So, we set

$$\frac{4k}{3} + \bar{t} + \varepsilon k = \frac{3k}{2} + \frac{\hat{p}}{2} + \varepsilon k \quad \Rightarrow \quad \bar{t} = \frac{k}{6} + \frac{\hat{p}}{2}$$

Applying this to (1.5), we have

$$\hat{p}(n-k) + \frac{n(\frac{k}{6}+\frac{\hat{p}}{2})}{2} \le nk - \frac{n^2}{2} + O(k) \quad \Rightarrow \quad \hat{p}\left(\frac{5n}{4} - k\right) \le \frac{11}{12}nk - \frac{n^2}{2} + O(k)$$

Therefore, $n$ is at most the maximum of $1.59k$ and

$$\frac{3k}{2}+\frac{11nk-6n^2+O(k)}{2(15n-12k)}+\varepsilon k \leq \frac{3k}{2}+\frac{11nk-6n^2}{2(15n-12k)}+2\varepsilon k. \tag{1.6}$$

Note that the right hand side of (1.6) is decreasing in $n$. So, assuming $n \geq 1.59k$, we conclude $n \leq \eta k + 2\varepsilon k$, where $\eta$ is a root of the equation

$$(2\eta-3)(15\eta-12) = 11\eta - 6\eta^2$$

in the interval $[1,3]$. As this quadratic equation simplifies to $9\eta^2 - 20\eta + 9 = 0$, we have

$$n \leq \left(\frac{10+\sqrt{19}}{9}+2\varepsilon\right)k$$

for sufficiently large $k$, as desired.

## 1.4 Extremal number of the messy path

We define $S^{(3)}_{n-1}$ to be the complete 3-uniform star on $n$ vertices; that is, $S^{(3)}_{n-1}$ is the collection of all triples on $n$ vertices that contain some fixed vertex. Füredi and Özkahya proved that for sufficiently large $n$, $ex(n,\mathscr{M}) = \binom{n-1}{2}$ and $S^{(3)}_{n-1}$ is the unique extremal hypergraph [FO11]. Here we observe that this result extends to smaller $n$.

Of course, the star is not the unique extremal hypergraph in $Ex(n,\mathscr{M})$ when $n$ is too small. For example, the complete 3-uniform hypergraph on $n$ vertices does not contain $\mathscr{M}$ when $n \leq 5$. Furthermore, if $X$ is a vertex set of size $n = 6$ then any collection $\mathscr{F}$ of 3-sets with the property that $e \in \mathscr{F}$ implies $X \setminus e \notin \mathscr{F}$ does not contain a copy of the messy path. The collection of all such hypergraphs is given explicitly in [PR17a]. Our main result here is that the star is the unique extremal hypergraph in $Ex(n,\mathscr{M})$ if $n \geq 7$.

In the proof we make use of hypergraph $F(a,2)$ which was introduced above. Recall that $F(a,2)$ refers to the 3-uniform hypergraph with vertex set $\{x_1,\ldots,x_a,y_1,y_2\}$ and edge set $\{x_iy_1y_2 : i \in [a]\}$. We refer to the vertices $x_1,\ldots,x_a$ as petals of the $F(a,2)$ and $y_1,y_2$ as the center.

*Remark* 1.4.1. A family $\mathscr{F} \subset \binom{X}{3}$ that does not contain a messy path has the property that if $A, B, C \in \mathscr{F}$ are distinct sets such that $|A \cup B \cup C| = 6$ then $A \cap B \neq \varnothing, A \cap C \neq \varnothing, B \cap C \neq \varnothing$. It is known that for $n$ sufficiently large, the size of such a family is at most $\binom{n-1}{2}$, with the unique extremal family being a star [FO11]. This extremal result is also known for other similar conditions on families of sets generalizing the Erdős-Ko-Rado Theorem [EKR61, FF83, Mub06, MV05].

**Theorem 1.4.2.**

$$ex(n, \mathscr{M}) = \begin{cases} \binom{n}{3} & \text{if } n \leq 5 \\ \binom{n-1}{2} & \text{if } n \geq 6. \end{cases}$$

*Furthermore, if $n \geq 7$ then $S^{(3)}_{n-1}$ is the unique hypergraph in $Ex(n, \mathscr{M})$.*

*Proof.* For $n \leq 5$, since a messy path has 6 vertices, $\binom{[n]}{3}$ is the unique extremal hypergraph. For $n = 6$, note that if there is a pair of disjoint triples, then any other edge would create a messy path. So we may assume the hypergraph is intersecting, and Proposition 1.6 in [PR17a] gives the desired result.

For $n \geq 7$, we proceed by induction. For the base case of $n = 7$, it suffices to show that the family is intersecting—then the result follows by Erdős-Ko-Rado bound. So, suppose there are two disjoint triples $e, f$. Note that any other triple contained in their union would create a messy path. Thus, all other triples must contain the last vertex. Note that if this vertex is incident with an triple intersecting $e$ in two vertices and another triple intersecting $f$ in two vertices, then this would also create a messy path. Thus, there are no triples contained among this vertex and for instance, $f$. This implies that there are at most $\binom{6}{2} - 3 + 2$ triples. Thus, the star is the unique extremal family by Erdős-Ko-Rado.

For $n \geq 8$, we assume that we have messy path-free family $\mathscr{H}$ on vertex set $V$ with at least $\binom{n-1}{2}$ triples. Note that if some vertex has degree at most $n - 2 = \binom{n-2}{1}$, then the remaining $n - 1$ vertices are messy path-free with at least $\binom{n-2}{2}$ triples, so by the inductive hypothesis, this is exactly $\binom{n-2}{2}$ triples in a complete star. Note that if any triple from the removed vertex does not contain the star center then we have a messy path. Thus, the extremal family would be a complete star having exactly $\binom{n-1}{2}$ triples.

So, it suffices to show that $\mathscr{H}$ has a vertex of degree at most $n-2$. Assume for the sake of contradiction that no such vertex exists.

Note that the average codegree in $\mathscr{H}$ satisfies

$$\frac{3|\mathscr{H}|}{\binom{n}{2}} \geq \frac{3\binom{n-1}{2}}{\binom{n}{2}} > 2.$$

So, we may assume $\mathscr{H}$ has a pair of vertices $u, v$ with co-degree at least 3.

We first consider the case where there exists two vertices $u, v$ of codegree at least 4. Consider any two petals $x, y$ of the resulting $F(a,2)$ with center $\{u, v\}$. Recall that, by assumption, both $x$ and $y$ have degree at least $n-1$. Note that every triple through $x$ or $y$ must intersect $u$ or $v$ or we have a messy path. Thus, each of $x, y$ has at least $n-2$ triples intersecting exactly one of $u$ or $v$, of which at least $n-4$ do not contain both $x, y$. Since $n-4 \geq 3$, we note that $x$ must have at least 2 triples intersecting say, without loss of generality, $u$ which do not contain $v$. Let these triples be $uxa, uxb$. Then, if $y$ is in a triple $vyc$ for $c$ not $u, x$, taking $vyc$, one of $uxa$ and $uxb$ (as to not include $c$), and $uvy$, we would have a messy path. Thus, all triples through $y$ except possibly $vyx$ must intersect $u$. Let another petal of this $F(a,2)$ be $z$. Repeat this argument with $x, z$ to get that $z$ has at least 2 triples through $u$ but not $v, x, y$. Finally, repeating this again with $z, y$ to get that all triples through $y$ must intersect $u$ except possibly $vyz$, we conclude that all triples through $y$ must contain $u$, so $y$ has degree at most $n-2$, a contradiction.

Otherwise all pairs of vertices have codegree at most 3. Now consider any petal $x$ of the $F(3,2)$ with center $\{u, v\}$ and recall that it has degree at least $n-1$. Note that at most one triple through $x$ can avoid $u, v$ in this case. Thus, $x$ has at least $n-3$ triples intersecting exactly one of $u$ or $v$. Then, we have that $\max(\deg(x,u), \deg(x,v)) \geq 1 + \frac{n-3}{2}$ where the 1 counts $uvx$. Then, $\frac{n-1}{2} \geq \frac{7}{2} > 3$, a contradiction.

Thus, there must exist a vertex of degree at most $n-2$ and by induction, the proof is complete. $\square$

## 1.5 Conclusion

We emphasize that we spent little effort optimizing for the second order terms (i.e., $\varepsilon$) in Theorems 2 and 3. One reason for this is that we suspect that incremental improvements on the bounds we prove here can be achieved with a bit more effort. In other words, we suspect that our results here do not give the correct asymptotics of these multicolor Ramsey numbers. The main barrier we see to significant improvement on our upper bounds is the proliferation of 2-cycles in the digraphs introduced in Sections 2.2 and 3.2 that can occur when $n$ is close to $k$. Indeed, all of the methods that we use in this work are based on pairs of vertices with arcs oriented in only one direction in these digraphs. It would be interesting to see methods that could handle these 2-cycles and thereby produce upper bounds on $r_k(\mathscr{H})$ that are dramatically closer to $k$.

There are a number of other 3-edge triple systems for which the multicolor Ramsey number is an interesting open question. These include the loose cycle $\mathscr{C} = \{abc, cde, afe\}$ and the hypergraph $\mathscr{F}_5 = \{abc, abd, cde\}$. The best known bounds for these multicolor Ramsey numbers are as follows. For the loose cycle,

$$k+5 \leq r_k(\mathscr{C}) \leq 3k \text{ for } k \geq 3,$$

analogous to the simple bounds for $\mathscr{L}$ and $\mathscr{M}$ [GR12]. For $\mathscr{F}_5$,

$$2^{ck} \leq r_k(F_5) \leq k! \text{ for } k \geq 4 \text{ and } c \text{ some positive constant,}$$

which resembles bounds for $K_4^{(3)} - e$ and $K_3$ [AGLM14]. The hypergraph $\mathscr{G} = \{abc, abd, bef\}$, which we dub the giraffe, was addressed in the masters thesis of the second author, who showed

$$k+1 \leq r_k(\mathscr{K}) \leq r_k(\mathscr{G}) \leq k+4,$$

where the kite $\mathscr{K} = \{abc, abd\}$ has $k+1 \leq r_k(\mathscr{K}) \leq k+3$ [AGLM14].

### 1.5.1 Acknowledgements

Chapter 1, in full, is a reprint of the material as it appears in the paper *On multicolor Ramsey numbers of triple system paths of length* 3. This paper was published in *SIAM Journal on Discrete Mathematics*, volume 37(3), 2023, pp. 1419–1435. It was co-authored by the dissertation author together with Tom Bohman.

# Chapter 2

# A note on the Erdős–Hajnal hypergraph Ramsey problem

We show that there is an absolute constant $c > 0$ such that the following holds. For every $n > 1$, there is a 5-uniform hypergraph on at least $2^{2^{cn^{1/4}}}$ vertices with independence number at most $n$, where every set of 6 vertices induces at most 3 edges. The double exponential growth rate for the number of vertices is sharp. By applying a stepping-up lemma established by the first two authors, analogous sharp results are proved for $k$-uniform hypergraphs. This answers the penultimate open case of a conjecture in Ramsey theory posed by Erdős and Hajnal in 1972.

## 2.1 Introduction

The Ramsey number $r_k(s,n)$ is the minimum integer $N$ such that for any red/blue coloring of the $k$-tuples of $[N] = \{1,2,\ldots,N\}$, there is either a set of $s$ integers with all of its $k$-tuples colored red, or a set of $n$ integers with all of its $k$-tuples colored blue. Estimating $r_k(s,n)$ is a fundamental problem in combinatorics and has been extensively studied since 1935. For graphs, classical results of Erdős [Erd47] and Erdős and Szekeres [ES35] imply that $2^{n/2} < r_2(n,n) < 2^{2n}$. While small improvements have been made in both the upper and lower bounds for $r_2(n,n)$ (see [Con09, Spe78]), the constant factors in the exponents have not changed over the last 75 years.

Unfortunately for 3-uniform hypergraphs, there is an exponential gap between the best known upper and lower bounds for $r_3(n,n)$. Namely, Erdős, Hajnal, and Rado [EHR65, ER52] showed that

$$2^{cn^2} < r_3(n,n) < 2^{2^{c'n}},$$

where $c$ and $c'$ are absolute constants. For $k \geq 4$, their results also imply an exponential gap between the lower and upper bounds for $r_k(n,n)$,

$$\mathrm{twr}_{k-1}(cn^2) < r_k(n,n) < \mathrm{twr}_k(c'n),$$

where the *tower function* is defined recursively as $\mathrm{twr}_1(x) = x$ and $\mathrm{twr}_{i+1} = 2^{\mathrm{twr}_i(x)}$. Determining the tower growth rate of $r_k(n,n)$ is one of the most central problems in extremal combinatorics. Erdős, Hajnal, and Rado conjectured that the upper bound is closer to the truth, namely $r_k(n,n) = \mathrm{twr}_k(\Theta(n))$, and Erdős offered a \$500 reward for a proof (see [Chu97]).

*Off-diagonal* Ramsey numbers $r_k(s,n)$ have also been extensively studied. Here, $k$ and $s$ are fixed constants and $n$ tends to infinity. It follows from well-known results that $r_2(s,n) = n^{\Theta(1)}$ (see [AKS80, Boh09, BK10, ER52] for the best known bounds), and for 3-uniform hypergraphs, $r_3(s,n) = 2^{n^{\Theta(1)}}$ (see [CFS10] for the best known bounds).

For $k > 3$, Erdős, Hajnal, and Rado showed that $r_k(s,n) \leq \mathrm{twr}_{k-1}(n^c)$ where $c = c(k,s)$, and Erdős and Hajnal conjectured that this bound is the correct tower growth rate. In [MS18], the first two authors verified the conjecture for $s \geq k+2$, and for the last case $s = k+1$, they showed that $r_k(k+1,n) \geq \mathrm{twr}_{k-2}(n^{c\log n})$. Hence, there remains an exponential gap between the best known lower and upper bounds for $r_k(k+1,n)$ for $k \geq 4$.

Due to our lack of understanding of $r_k(k+1,n)$, Erdős and Hajnal in [EH72] introduced the following more general function (their notation was different).

**Definition 2.1.1.** For integers $2 \leq k < n$ and $2 \leq t \leq k+1$, let $r_k(k+1,t;n)$ be the minimum $N$ such that for every red/blue coloring of the $k$-tuples of $[N]$, there is a set of $k+1$ integers with at least $t$ of its $k$-tuples colored red, or a set of $n$ integers with all of its $k$-tuples colored blue.

Clearly $r_k(k+1,1;n) = n$ and $r_k(k+1,k+1;n) = r_k(k+1,n)$. For each $t \in \{2,\dots,k\}$, Erdős and Hajnal [EH72] showed that $r_k(k+1,t;n) < \text{twr}_{t-1}(n^{\Theta(1)})$ and conjectured that

$$r_k(k+1,t;n) = \text{twr}_{t-1}(n^{\Theta(1)}). \tag{2.1}$$

This is known to be true for $k \leq 3$ and for $t \leq 3$ [EH72]. When $k \geq 5$, the first two authors [MS20] verified (2.1) for all $3 \leq t \leq k-2$. Our main result verifies (2.1) for $t = k-1$, which is one of the last two remaining cases.

**Theorem 2.1.2.** *For $k \geq 4$, we have $r_k(k+1,k-1;n) = \text{twr}_{k-2}(n^{\Theta(1)})$.*

This significantly improves the previous best known lower bound for $r_k(k+1,k-1;n)$, which was one exponential less than above (see [MS20]). This also immediately implies the following new lower bound for $r_k(k+1,k;n)$, which is now one exponential off from the upper bound obtained by Erdős and Hajnal.

**Corollary 2.1.3.** *For $k \geq 4$, we have $r_k(k+1,k;n) > \text{twr}_{k-2}(n^{\Theta(1)})$.*

Finally, let us point out that Erdős and Hajnal conjectured that the tower growth rate for both $r_k(k+1,k;n)$ and the classical Ramsey number $r_k(k+1,n)$ are the same. Thus, verifying (2.1) for $r_k(k+1,k;n)$ would determine the tower height for $r_k(k+1,n)$.

We develop several crucial new ingredients to the stepping up method in our construction, for example, part (1) of Lemma 2.2.3, and on page 8, analyzing sequences of local maxima. It is plausible that these new ideas can be further enhanced to determine the tower height of $r_k(k+1,n)$.

## 2.2 Proof of Theorem 2.1.2

In [MS18], the first two authors proved the following.

**Theorem 2.2.1** (Theorem 7 in [MS18])**.** *For $k \geq 6$ and $t \geq 5$, we have*

$$r_k(k+1,t;2kn) > 2^{r_{k-1}(k,t-1;n)-1}.$$

In what follows, we will prove the following theorem. Together with Theorem 2.2.1, Theorem 2.1.2 quickly follows.

**Theorem 2.2.2.** *There is an absolute constant $c > 0$ such that $r_5(6,4;n) > 2^{2^{cn^{1/4}}}$.*

### 2.2.1 A double exponential lower bound for $r_5(6,4;n)$

In this section, we begin with a graph coloring with certain properties which we will later use to define a two-coloring of the edges of a 5-uniform hypergraph.

**Lemma 2.2.3.** *For $n \geq 6$, there is an absolute constant $c > 0$ such that the following holds. There exists a red/blue coloring $\phi$ of the pairs of $\{0,1,\ldots,\lfloor 2^{cn} \rfloor\}$ such that:*

1. *There are no 3 disjoint n-sets $A,B,C \subset \{0,1,\ldots,\lfloor 2^{cn} \rfloor\}$ with the property that there is a bijection $f : B \to C$ such that for any $a \in A, b \in B$, at least one of $\phi(a,b) = red$ or $\phi(a, f(b)) = blue$ occurs.*
2. *There is no n-set $A \subset \{0,1,\ldots,\lfloor 2^{cn} \rfloor\}$ such that every 4-tuple $a_i,a_j,a_k,a_\ell \in A$ with $a_i < a_j < a_k < a_\ell$ avoids the pattern:*

$$\phi(a_i,a_j) = \phi(a_j,a_k) = \phi(a_j,a_\ell) = red, \qquad \phi(a_i,a_k) = \phi(a_i,a_\ell) = \phi(a_k,a_\ell) = blue$$

*Proof.* Set $N = \lfloor 2^{cn} \rfloor$, where $c$ is a sufficiently small constant that will be determined later. Consider a random 2-coloring of the unordered pairs of $\{0,1,\ldots,N-1\}$ where each pair is assigned red or blue with equal probability independent of all other pairs. Then, the expected number of $A,B,C$ as in part 1 is at most

$$\binom{N}{n}^3 n! \left(\frac{3}{4}\right)^{n^2} < \frac{1}{3},$$

where the inequality holds by taking $c$ sufficiently small. This is since we pick each of the $n$-sets, one of $n!$ possible bijections from $B$ to $C$, and then there is a $\frac{3}{4}$ probability that we have the desired color pattern for each pair of $a \in A, b \in B$.

We call a 4-tuple $a_i, a_j, a_k, a_\ell \in \{0, 1, \ldots, N-1\}$ with $a_i < a_j < a_k < a_\ell$ *bad* if

$$\phi(a_i, a_j) = \phi(a_j, a_k) = \phi(a_j, a_\ell) = \text{red}, \qquad \phi(a_i, a_k) = \phi(a_i, a_\ell) = \phi(a_k, a_\ell) = \text{blue}$$

and *good* otherwise. The probability that such a fixed 4-tuple is bad is $\frac{1}{2^6} = \frac{1}{64}$ and thus the probability that such a fixed 4-tuple is good is $\frac{63}{64}$. Now consider some fixed $n$-set $A \subset \{0, 1, \ldots, N-1\}$. We estimate the probability that $A$ contains no bad 4-tuple. Note that there exists a partial Steiner $(n, 4, 2)$-system $S$ on $A$, i.e. a 4-uniform hypergraph on the $n$-vertex set $A$ with the property that every pair of vertices is contained in at most one 4-tuple, with at least $c'n^2$ edges where $c' > 0$ is some constant (e.g. see [EH63]). Then, the probability that a 4-tuple in $A$ is good is at most the probability that every 4-tuple in $S$ is good. Since 4-tuples in $S$ are independent as no two 4-tuples have more than one vertex is common, the probability that every 4-tuple in $S$ is a good 4-tuple is at most $\left(\frac{63}{64}\right)^{c'n^2}$. Therefore, the expected number of $n$-sets $A$ with only good 4-tuples is at most

$$\binom{N}{n} \left(\frac{63}{64}\right)^{c'n^2} < \frac{1}{3},$$

again where we take $c$ sufficiently small. Thus, by Markov's inequality and the union bound, we conclude that there is a 2-coloring $\phi$ with the desired properties. □

We will use this lemma to produce a coloring of a 5-uniform hypergraph. Given some natural number $N$, let $V = \{0, 1, \ldots, 2^N - 1\}$. Then for $v \in V$, we write $v = \sum_{i=0}^{N-1} v(i) 2^i$ where $v(i) \in \{0, 1\}$ for each $i$. For any $u \neq v$, we then let $\delta(u, v)$ denote the largest $i \in \{0, 1, \ldots, N-1\}$ such that $u(i) \neq v(i)$. We then have the following properties.

**Property I**: For every triple $u < v < w, \delta(u, v) \neq \delta(v, w)$.

**Property II**: For $v_1 < \cdots < v_r, \delta(v_1, v_r) = \max_{1 \leq j \leq r-1} \delta(v_j, v_{j+1})$.

From Properties I and II, we also derive the following.

**Property III**: For every 4-tuple $v_1 < \cdots < v_4$, if $\delta(v_1, v_2) > \delta(v_2, v_3)$, then $\delta(v_1, v_2) \neq \delta(v_3, v_4)$. Note that if $\delta(v_1, v_2) < \delta(v_2, v_3)$, it is possible that $\delta(v_1, v_2) = \delta(v_3, v_4)$.

**Property IV**: For $v_1 < \cdots < v_r$, set $\delta_j = \delta(v_j, v_{j+1})$ for $j \in [r-1]$ and suppose that

$\delta_1, \ldots, \delta_{r-1}$ forms a monotone sequence. Then for every subset of $k$ vertices $v_{i_1}, v_{i_2}, \ldots, v_{i_k}$ where $v_{i_1} < \cdots < v_{i_k}$, $\delta(v_{i_1}, v_{i_2}), \delta(v_{i_2}, v_{i_3}), \ldots, \delta(v_{i_{k-1}}, v_{i_k})$ forms a monotone sequence. Moreover for every subset of $k-1$ such $\delta_j$'s, i.e. $\delta_{j_1}, \delta_{j_2}, \ldots, \delta_{j_{k-1}}$, there are $k$ vertices $v_{i_1}, \ldots, v_{i_k}$ such that $\delta(v_{i_t}, v_{i_{t+1}}) = \delta_{j_t}$.

We now turn to the coloring of a 5-uniform hypergraph. Let $c > 0$ be the constant given by Lemma 2.2.3 and let $U = \{0, 1, \ldots, \lfloor 2^{cn} \rfloor\}$ and $\phi : \binom{U}{2} \to \{\text{red}, \text{blue}\}$ be a 2-coloring of the pairs of $U$ satisfying the properties given in the lemma. Now let $N = 2^{\lfloor 2^{cn} \rfloor}$ and let $V = \{0, 1, \ldots, N-1\}$. In the following, we will use the coloring $\phi$ to define a red/blue coloring $\chi : \binom{V}{5} \to \{\text{red}, \text{blue}\}$ of the 5-tuples of $V$ such that $\chi$ produces at most 3 red edges among any 6 vertices and $\chi$ does not produce a blue copy of $K^{(5)}_{128n^4}$. This would imply that $r_5(6,4;n) > 2^{2^{c'n^{1/4}}}$ for some constant $c' > 0$.

For $v_1, \ldots, v_5 \in V$ with $v_1 < v_2 < \cdots < v_5$, let $\delta_i = \delta(v_i, v_{i+1})$. We set $\chi(v_1, \ldots, v_5) = \text{red}$ if:

1. We have that $\delta_1, \delta_2, \delta_3, \delta_4$ are monotone and form a bad 4-tuple, that is, if $\delta_1 < \delta_2 < \delta_3 < \delta_4$ then:

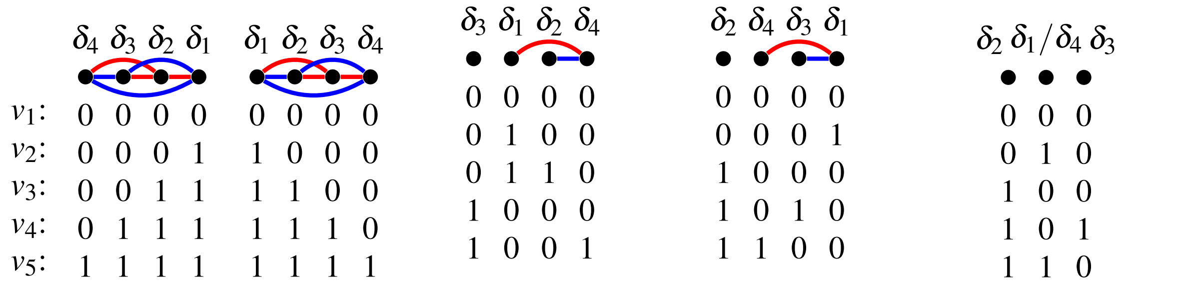
   $$\phi(\delta_1, \delta_2) = \phi(\delta_2, \delta_3) = \phi(\delta_2, \delta_4) = \text{red}, \qquad \phi(\delta_1, \delta_3) = \phi(\delta_1, \delta_4) = \phi(\delta_3, \delta_4) = \text{blue},$$

   and if $\delta_1 > \delta_2 > \delta_3 > \delta_4$ then:

   $$\phi(\delta_4, \delta_3) = \phi(\delta_3, \delta_2) = \phi(\delta_3, \delta_1) = \text{red}, \qquad \phi(\delta_4, \delta_2) = \phi(\delta_4, \delta_1) = \phi(\delta_2, \delta_1) = \text{blue}.$$

2. We have that $\delta_1 > \delta_2 < \delta_3 > \delta_4$, where $\delta_1, \delta_2, \delta_3, \delta_4$ are all distinct with $\delta_1 < \delta_3, \delta_2 > \delta_4$ and $\phi(\delta_1, \delta_4) = \text{red}$, $\phi(\delta_2, \delta_4) = \text{blue}$. The ordering can also be expressed as $\delta_3 > \delta_1 > \delta_2 > \delta_4$.

3. We have that $\delta_1 < \delta_2 > \delta_3 < \delta_4$, where $\delta_1, \delta_2, \delta_3, \delta_4$ are all distinct with $\delta_1 < \delta_3, \delta_2 > \delta_4$ and $\phi(\delta_1, \delta_4) = \text{red}$, $\phi(\delta_1, \delta_3) = \text{blue}$. The ordering can also be expressed as $\delta_2 > \delta_4 > \delta_3 > \delta_1$.




**(a)** Monotone **(b)** $\delta_3 > \delta_1 > \delta_2 > \delta_4$ **(c)** $\delta_2 > \delta_4 > \delta_3 > \delta_1$ **(d)** $\delta_2 > \delta_1 = \delta_4 > \delta_3$

**Figure 2.1.** Examples of $v_1 < v_2 < v_3 < v_4 < v_5$ and $\delta_i = \delta(v_i, v_{i+1})$ for $i \in [4]$ such that $\chi(v_1, \ldots, v_5)$ is red. Each $v_i$ is represented in binary with the left-most entry corresponding to the most significant bit.

4. We have that $\delta_1 < \delta_2 > \delta_3 < \delta_4$ and $\delta_1 = \delta_4$. In other words, $\delta_2 > \delta_1 = \delta_4 > \delta_3$.

Otherwise $\chi(v_1, \ldots, v_5) = \text{blue}$.

Assume for the sake of contradiction that there are at least 4 red edges among some 6 vertices. Let these vertices be $v_1, \ldots, v_6$ where $v_1 < v_2 < \cdots < v_6$ and let $\delta_i = \delta(v_i, v_{i+1})$. Let $e_i = \{v_1, \ldots, v_6\} \setminus \{v_i\}$. Let $\delta(e_i)$ be the resulting sequence of $\delta$'s. In particular, for $i = 1$, $\delta(e_1) = (\delta_2, \delta_3, \delta_4, \delta_5)$. For $2 \leq i \leq 5$, $\delta(e_i) = (\delta_1, \ldots, \delta(v_{i-1}, v_{i+1}), \ldots, \delta_5)$. For $i = 6$, $\delta(e_6) = (\delta_1, \delta_2, \delta_3, \delta_4)$. In the following we will often use that if $2 \leq i \leq 5$, then $\delta(v_{i-1}, v_{i+1}) = \max(\delta_{i-1}, \delta_i)$ by Property II.

For convenience, if inequalities are known between consecutive $\delta$'s, this will be indicated in the sequence by replacing the comma with the respective sign. For instance, assume that $\delta_1 < \delta_2 > \delta_3 < \delta_4 > \delta_5$. Then since $\delta(e_1) = (\delta_2, \delta_3, \delta_4, \delta_5)$ has $\delta_2 > \delta_3 < \delta_4 > \delta_5$, we will write

$$\delta(e_1) = (\delta_2 > \delta_3 < \delta_4 > \delta_5).$$

Similarly, if not all inequalities are known, as in $\delta(e_3)$, we write,

$$\delta(e_3) = (\delta_1 < \delta_2 \ , \ \delta_4 > \delta_5).$$

Now we will consider cases depending on the ordering of $\delta_1, \ldots, \delta_5$, and we will further split into subcases by taking an ordering and reversing it. There are 16 possible orderings so we will have 8 cases in what follows.

**Case 1a**: Suppose $\delta_1 > \delta_2 < \delta_3 > \delta_4 < \delta_5$. This implies that

$$\delta(e_1) = (\delta_2 < \delta_3 > \delta_4 < \delta_5),$$

$$\delta(e_2) = \delta(e_3) = (\delta_1 \ , \ \delta_3 > \delta_4 < \delta_5),$$

$$\delta(e_4) = \delta(e_5) = (\delta_1 > \delta_2 < \delta_3 \ , \ \delta_5),$$

$$\delta(e_6) = (\delta_1 > \delta_2 < \delta_3 > \delta_4).$$

In particular, note that at least one of $e_4, e_5, e_6$ must be red so we must have that $\delta_1 < \delta_3$ and $\delta_2 > \delta_4$. However, since $\delta_1 > \delta_2 > \delta_4$, note that $e_1$ is only red if $\delta_2 = \delta_5$ and similarly $e_2, e_3$ are only red if $\delta_1 = \delta_5$. Since these cannot happen simultaneously, there is at least one blue edge among these three edges. Thus, we must have that $e_4, e_5$ are also red to avoid having three blue edges, making $\delta_2 > \delta_5$ (and $\delta_3 > \delta_5$). However, then $\delta_1 > \delta_2 > \delta_5$ so none of $e_1, e_2, e_3$ are red and thus there are at most 3 red edges.

**Case 1b**: Suppose $\delta_1 < \delta_2 > \delta_3 < \delta_4 > \delta_5$. This implies that

$$\delta(e_1) = \delta(e_2) = (\delta_2 > \delta_3 < \delta_4 > \delta_5),$$

$$\delta(e_3) = \delta(e_4) = (\delta_1 < \delta_2 \ , \ \delta_4 > \delta_5),$$

$$\delta(e_5) = \delta(e_6) = (\delta_1 < \delta_2 > \delta_3 < \delta_4).$$

Note that $e_3, e_4$ are blue so we must have that all of $e_1, e_2, e_5, e_6$ are red. If $e_5, e_6$ are red, then regardless of which rule applies, $\delta_2 > \delta_4$ and thus $e_1, e_2$ are blue, so there are at most 2 red edges.

**Case 2a**: Suppose $\delta_1 > \delta_2 > \delta_3 < \delta_4 > \delta_5$. This implies that

$$\delta(e_1) = (\delta_2 > \delta_3 < \delta_4 > \delta_5),$$

$$\delta(e_2) = (\delta_1 > \delta_3 < \delta_4 > \delta_5),$$

$$\delta(e_3) = \delta(e_4) = (\delta_1 > \delta_2 \ , \ \delta_4 > \delta_5),$$

$$\delta(e_5) = \delta(e_6) = (\delta_1 > \delta_2 > \delta_3 < \delta_4).$$

Note that $e_5, e_6$ are blue so that all of $e_1, \ldots, e_4$ are red. Since $e_1$ is red, we must have that $\delta_2 < \delta_4$, so $\delta(e_i)$ are ordered as in the second condition for red edges for all $i \in [4]$. Thus, $e_1$ implies that $\phi(\delta_2, \delta_5) =$ red while $e_3$ implies that $\phi(\delta_2, \delta_5) =$ blue, a contradiction.

**Case 2b**: Suppose $\delta_1 < \delta_2 < \delta_3 > \delta_4 < \delta_5$. This implies that

$$\delta(e_1) = \delta(e_2) = (\delta_2 < \delta_3 > \delta_4 < \delta_5),$$

$$\delta(e_3) = (\delta_1 < \delta_3 > \delta_4 < \delta_5),$$

$$\delta(e_4) = \delta(e_5) = (\delta_1 < \delta_2 < \delta_3 \ , \ \delta_5),$$

$$\delta(e_6) = (\delta_1 < \delta_2 < \delta_3 > \delta_4).$$

Since $e_6$ is blue, in order to have at least 4 red edges, we must have that $e_4, e_5$ are red. Thus $\delta_3 < \delta_5$. However, then for $e_1, e_2$ to be red, we must have that $\delta_2 = \delta_5$, which is impossible since $\delta_2 < \delta_5$. Thus, there are at most 3 red edges here.

**Case 3a**: Suppose $\delta_1 > \delta_2 < \delta_3 > \delta_4 > \delta_5$. This implies that

$$\delta(e_1) = (\delta_2 < \delta_3 > \delta_4 > \delta_5),$$

$$\delta(e_2) = \delta(e_3) = (\delta_1 \ , \ \delta_3 > \delta_4 > \delta_5),$$

$$\delta(e_4) = (\delta_1 > \delta_2 < \delta_3 > \delta_5),$$

$$\delta(e_5) = \delta(e_6) = (\delta_1 > \delta_2 < \delta_3 > \delta_4).$$

Since $e_1$ is blue, we must have that $e_5, e_6$ are red and thus $\delta_1 < \delta_3$. However, we also must have $e_2, e_3$ are red and thus $\delta_1 > \delta_3$, a contradiction.

**Case 3b**: Suppose $\delta_1 < \delta_2 > \delta_3 < \delta_4 < \delta_5$. This implies that

$$\delta(e_1) = \delta(e_2) = (\delta_2 > \delta_3 < \delta_4 < \delta_5),$$

$$\delta(e_3) = \delta(e_4) = (\delta_1 < \delta_2 \ , \ \delta_4 < \delta_5),$$

$$\delta(e_5) = (\delta_1 < \delta_2 > \delta_3 < \delta_5),$$

$$\delta(e_6) = (\delta_1 < \delta_2 > \delta_3 < \delta_4).$$

Since $e_1, e_2$ are blue, we must have that the remaining edges are red. If $\delta_2 < \delta_4$, then $e_6$ is blue. Otherwise $\delta_2 > \delta_4$. First if $\delta_1 = \delta_4$ then $e_3, e_4$ are blue. Thus, for $e_6$ to be red, we have that $\delta_1 < \delta_3$, which implies that $\delta_1 < \delta_4 < \delta_5$. From $e_3$ being red, we find that $\delta_2 > \delta_5$ as well. We then have that $\phi(\delta_1, \delta_4) = \text{red}$ from $e_6$ while $\phi(\delta_1, \delta_4) = \text{blue}$ from $e_3$, a contradiction.

**Case 4a**: Suppose $\delta_1 > \delta_2 < \delta_3 < \delta_4 > \delta_5$. This implies that

$$\delta(e_1) = (\delta_2 < \delta_3 < \delta_4 > \delta_5),$$

$$\delta(e_5) = \delta(e_6) = (\delta_1 > \delta_2 < \delta_3 < \delta_4).$$

so we have at least 3 blue edges.

**Case 4b**: Suppose $\delta_1 < \delta_2 > \delta_3 > \delta_4 < \delta_5$. This implies that

$$\delta(e_1) = \delta(e_2) = (\delta_2 > \delta_3 > \delta_4 < \delta_5),$$

$$\delta(e_6) = (\delta_1 < \delta_2 > \delta_3 > \delta_4).$$

so we have at least 3 blue edges.

**Case 5**: Suppose $\delta_1 > \delta_2 < \delta_3 < \delta_4 < \delta_5$ or $\delta_1 < \delta_2 > \delta_3 > \delta_4 > \delta_5$. In the first case, each of $\delta(e_4), \delta(e_5), \delta(e_6)$ is in the form $\delta_1 > \delta_2 < \delta_i < \delta_j$ where $i, j \in \{3,4,5\}$, so these are blue. In the second case, each of $\delta(e_4), \delta(e_5), \delta(e_6)$ is in the form $\delta_1 < \delta_2 > \delta_i > \delta_j$ where $i, j \in \{3,4,5\}$, so these are blue.

**Case 6**: Suppose $\delta_1 > \delta_2 > \delta_3 < \delta_4 < \delta_5$ or $\delta_1 < \delta_2 < \delta_3 > \delta_4 > \delta_5$. In the first case,

$$\delta(e_1) = (\delta_2 > \delta_3 < \delta_4 < \delta_5),$$

$$\delta(e_2) = (\delta_1 > \delta_3 < \delta_4 < \delta_5),$$

$$\delta(e_6) = (\delta_1 > \delta_2 > \delta_3 < \delta_4).$$

so there are at least 3 blue edges. In the second case, $\delta(e_1), \delta(e_2)$ are both $\delta_2 < \delta_3 > \delta_4 > \delta_5$ and thus blue. Similarly, $\delta(e_6) = \delta_1 < \delta_2 < \delta_3 > \delta_4$, so there are at least 3 blue edges.

**Case 7**: Suppose $\delta_1 > \delta_2 > \delta_3 > \delta_4 < \delta_5$ or $\delta_1 < \delta_2 < \delta_3 < \delta_4 > \delta_5$. In the first case, each of $\delta(e_1), \delta(e_2), \delta(e_3)$ is in the form $\delta_i > \delta_j > \delta_4 < \delta_5$ for $i, j \in [3]$ and thus blue. In the second case, each of $\delta(e_1), \delta(e_2), \delta(e_3)$ is in the form $\delta_i < \delta_j < \delta_4 > \delta_5$ for $i, j \in [3]$ and thus blue.

**Case 8a**: Suppose $\delta_1 > \delta_2 > \delta_3 > \delta_4 > \delta_5$. This implies that

$$\delta(e_1) = (\delta_2 > \delta_3 > \delta_4 > \delta_5),$$
$$\delta(e_2) = (\delta_1 > \delta_3 > \delta_4 > \delta_5),$$
$$\delta(e_3) = (\delta_1 > \delta_2 > \delta_4 > \delta_5),$$
$$\delta(e_4) = (\delta_1 > \delta_2 > \delta_3 > \delta_5),$$
$$\delta(e_5) = \delta(e_6) = (\delta_1 > \delta_2 > \delta_3 > \delta_4).$$

First if $e_5, e_6$ are red, then $\phi(\delta_4, \delta_1) =$ blue implies that $e_2, e_3$ are blue, and $\phi(\delta_4, \delta_2) =$ blue implies that $e_1$ is blue, a contradiction. Thus, $e_5, e_6$ are blue and $e_1$ must be red but then $\phi(\delta_5, \delta_3) =$ blue implies that $e_4$ is blue, a contradiction.

**Case 8b**: Suppose $\delta_1 < \delta_2 < \delta_3 < \delta_4 < \delta_5$. This implies that

$$\delta(e_1) = \delta(e_2) = (\delta_2 < \delta_3 < \delta_4 < \delta_5),$$
$$\delta(e_3) = (\delta_1 < \delta_3 < \delta_4 < \delta_5),$$
$$\delta(e_4) = (\delta_1 < \delta_2 < \delta_4 < \delta_5),$$
$$\delta(e_5) = (\delta_1 < \delta_2 < \delta_3 < \delta_5),$$
$$\delta(e_6) = (\delta_1 < \delta_2 < \delta_3 < \delta_4).$$

If $e_1, e_2$ are red, then $\phi(\delta_2, \delta_5) =$ blue implies that $e_4, e_5$ are blue and $\phi(\delta_2, \delta_4) =$ blue implies that $e_6$ is blue, a contradiction. Thus, $e_1, e_2$ are blue and $e_6$ must be red but then $\phi(\delta_1, \delta_3) =$ blue implies that $e_3$ is blue, a contradiction.

Thus, for every 6 vertices in $V = \{0, 1, \ldots, 2^{\lfloor 2^{cn} \rfloor} - 1\}$, $\chi$ produces at most 3 red edges among them.

Now, we show that there is no blue $K^{(5)}_{128n^4}$ in coloring $\chi$. We first make the following definitions. Given a sequence $\{a_i\}_{i=1}^{r} \subseteq \mathbb{R}$ and $j \in \{2, \ldots, r-1\}$, we say that $a_j$ is a *local minimum* if $a_{j-1} > a_j < a_{j+1}$, a *local maximum* if $a_{j-1} < a_j > a_{j+1}$, and a *local extremum* if it is either a local minimum or local maximum. In particular, when looking at some set of vertices $\{v_1, \ldots, v_s\}$ where $v_1 < v_2 < \cdots < v_s$ and considering the sequence $\{\delta(v_i, v_{i+1})\}_{i=1}^{s-1}$, by Property I, $\delta(v_j, v_{j+1}) \neq \delta(v_{j+1}, v_{j+2})$ for every $j$, so every nonmonotone sequence will have local extrema.

Set $m = 128n^4$ and consider vertices $v_1, \ldots, v_m \in V$ such that $v_1 < v_2 < \cdots < v_m$. Assume for the sake of contradiction that these $m$ vertices correspond to a blue clique in the coloring $\chi$. Again, let $\delta_i = \delta(v_i, v_{i+1})$. We first note the following lemma.

**Lemma 2.2.4.** *There is no monotone subsequence $\{\delta_{k_\ell}\}_{\ell=1}^{n} \subset \{\delta_i\}_{i=1}^{m-1}$ such that the following holds: for any $a, b, c, d \in [n]$ with $a < b < c < d$, there exists $u_1, u_2, u_3, u_4, u_5 \subset \{v_1, \ldots, v_m\}$ such that $\delta(u_1, \ldots, u_5) = \{\delta_{k_a}, \delta_{k_b}, \delta_{k_c}, \delta_{k_d}\}$.*

*Proof.* Indeed, if such a monotone subsequence existed, then as $\chi(u_1, \ldots, u_5) = \text{blue}$, we have that $\{\delta_{k_\ell}\}_{\ell=1}^{n}$ would form an $n$-set with no bad 4-tuple in the graph coloring $\phi$, a contradiction. □

From this, we note that there is no integer $j \in [m-n+1]$ such that the sequence $\{\delta_i\}_{i=j}^{j+n-1}$ is monotone. Otherwise, by Property IV, we have that for any length 4 subsequence $\{\delta_{i_1}, \delta_{i_2}, \delta_{i_3}, \delta_{i_4}\} \subset \{\delta_i\}_{i=j}^{j+n-1}$, there is a 5-tuple $e \subset \{v_1, \ldots, v_m\}$ such that $\delta(e)$ corresponds to this monotone sequence. From here, we apply Lemma 2.2.4 to get a contradiction. Thus, we can find a sequence of consecutive local extrema and from this extract a sequence of local maxima $\delta_{i_1}, \ldots, \delta_{i_{32n^3}}$.

We now restrict our attention to this sequence of local maxima $(\delta_{i_1}, \ldots, \delta_{i_{32n^3}})$. Note that any two local maxima are distinct: assume for the sake of contradiction that we have maxima $\delta_{i_j} = \delta_{i_k}$ where $j < k$. First consider if there is no $\delta_\ell$ for $i_j < \ell < i_k$ such that $\delta_\ell > \delta_{i_j} = \delta_{i_k}$. Then, $\delta(v_{i_j}, v_{i_k}) = \delta_{i_j} = \delta_{i_k} = \delta(v_{i_k}, v_{i_k+1})$, a contradiction of Property I. Otherwise, there exists $i_j < \ell < i_k$ such that $\delta_\ell > \delta_{i_j} = \delta_{i_k}$. By letting $\ell$ correspond to the maximum $\delta_\ell$ in this range, we have

$$\delta(v_{i_j}, v_{i_j+1}, v_{i_k-1}, v_{i_k}, v_{i_k+1}) = (\delta_{i_j} < \delta_\ell > \delta_{i_k-1} < \delta_{i_k}),$$

which implies that $\chi(v_{i_j}, v_{i_j+1}, v_{i_k-1}, v_{i_k}, v_{i_k+1}) = \text{red}$ as $\delta_{i_j} = \delta_{i_k}$, contradiction.

Moreover, there is no $j \in [32n^3 - n + 1]$ such that the sequence $\{\delta_{i_k}\}_{k=j}^{j+n-1}$ is monotone. If there is such $j$ and the sequence is increasing, for any $a, b, c, d \in \{j, j+1, \ldots, j+n-1\}$ with $a < b < c < d$, then

$$\delta(v_{i_a}, v_{i_a+1}, v_{i_b+1}, v_{i_c+1}, v_{i_d+1}) = (\delta_{i_a} < \delta_{i_b} < \delta_{i_c} < \delta_{i_d}).$$

This follows by Property II; in particular, if there exists $\ell$ such that $i_a + 1 \leq \ell < i_b + 1$ and $\delta_\ell > \delta_{i_b}$, then there must exist some greater local maxima between $\delta_{i_a}$ and $\delta_{i_b}$, a contradiction of the monotonicity of $\{\delta_{i_k}\}_{k=j}^{j+n-1}$, as these are consecutive local maxima. Thus, by Lemma 2.2.4, we have a contradiction.

Similarly, if the sequence is decreasing, consider any $a, b, c, d \in \{j, j+1, \ldots, j+n-1\}$ with $a < b < c < d$. Then

$$\delta(v_{i_a}, v_{i_b}, v_{i_c}, v_{i_d}, v_{i_d+1}) = (\delta_{i_a} > \delta_{i_b} > \delta_{i_c} > \delta_{i_d}).$$

As with the above, we apply Lemma 2.2.4 to derive a contradiction.

Thus, within the sequence $(\delta_{i_1}, \delta_{i_2}, \ldots, \delta_{i_{32n^3}})$, we can find a subsequence of consecutive local extrema $\delta_{j_1}, \ldots, \delta_{j_{16n^2}}$, where $\delta_{j_1}, \delta_{j_3}, \ldots, \delta_{j_{16n^2-1}}$ are local maxima and $\delta_{j_2}, \delta_{j_4}, \ldots, \delta_{j_{16n^2}}$ are local minima (with respect to the sequence $\delta_{i_1}, \delta_{i_2}, \ldots, \delta_{i_{32n^3}}$).

We now claim that there exists $k \in \{4n+1, 4n+2, \ldots, 16n^2 - 4n\}$ such that $\delta_{j_\ell} < \delta_{j_k}$ if $k - 4n \leq \ell \leq k + 4n$ and $\ell \neq k$. Assume for the sake of contradiction that this is not the case. We then recursively build the following sets $S_r, T_r$. Start with $S_0 = T_0 = \varnothing, \sigma_0 = 0, \tau_0 = 16n^2 + 1$. At each step $r$,

1. $\sigma_r = 0$ if $S_r$ is empty and $\sigma_r = \max(S_r)$ otherwise. Similarly, $\tau_r = 16n^2 + 1$ if $T_r$ is empty and $\tau_r = \min(T_r)$ otherwise.

2. If $s \in S_r$ and $s < \ell < \tau_r$, then $\delta_{j_s} > \delta_{j_\ell}$. Similarly if $t \in T_r$ and $\sigma_r < \ell < t$, then $\delta_{j_t} > \delta_{j_\ell}$.

3. $|S_r| + |T_r| = r$ and $\tau_r - \sigma_r \geq 16n^2 - 4nr$.

Note that these properties hold for $r = 0$ by definition. Now assume that we have $S_r, T_r, \sigma_r, \tau_r$ satisfying the desired properties for some $r < 2n$. Note that by the properties, we have that

$$\tau_r - \sigma_r \geq 16n^2 - 4nr > 16n^2 - 8n^2 \geq 8n^2 > 0.$$

Consider $\sigma_r < k < \tau_r$ such that $\delta_{j_k} = \max_{\sigma_r < \ell < \tau_r} \delta_{j_\ell}$. If $k - \sigma_r > 4n$ and $\tau_r - k > 4n$, then $k$ would satisfy that $\delta_{j_\ell} < \delta_{j_k}$ if $k - 4n \leq \ell \leq k + 4n$ and $\ell \neq k$, a contradiction. Now if $k - \sigma_r \leq 4n$, set

$$S_{r+1} = S_r \cup \{k\}, \quad T_{r+1} = T_r, \quad \sigma_{r+1} = k, \quad \tau_{r+1} = \tau_r.$$

Then, the first property holds by definition. The second property holds for every $s \in S_r, t \in T_r$ by assumption, and it holds for $k \in S_{r+1}$ since $\delta_{j_k} = \max_{\sigma_r < \ell < \tau_r} \delta_{j_\ell}$. The first part of the third property clearly holds and

$$\tau_{r+1} - \sigma_{r+1} = \tau_r - k \geq \tau_r - \sigma_r - 4n \geq 16n^2 - 4n(r+1).$$

Otherwise if $\tau_r - k \leq 4n$, set

$$S_{r+1} = S_r, \quad T_{r+1} = T_r \cup \{k\}, \quad \sigma_{r+1} = \sigma_r, \quad \tau_{r+1} = k.$$

By the same reasoning, the three properties hold as desired. Thus, we can construct these sets while $r \leq 2n$.

Now, consider $S_{2n}, T_{2n}$. Since $|S_{2n}| + |T_{2n}| = 2n$, at least one of these sets has size at least $n$. If $|S_{2n}| \geq n$, consider $\{s_1, \ldots, s_n\} \subseteq S_{2n}$ where $i < j \Rightarrow s_i < s_j$. Then, since $\min(T_{2n}) > \max(S_{2n})$ by Property 3 and 1, by Property 2 we have

$$\delta_{j_{s_1}} > \delta_{j_{s_2}} > \cdots > \delta_{j_{s_n}}.$$

In particular, Property 2 implies that for $a, b, c, d \in [n]$ and $a < b < c < d$,

$$\delta(v_{j_{s_a}}, v_{j_{s_b}}, v_{j_{s_c}}, v_{j_{s_d}}, v_{j_{s_d}+1}) = (\delta_{j_{s_a}} > \delta_{j_{s_b}} > \delta_{j_{s_c}} > \delta_{j_{s_d}}),$$

and thus, by Lemma 2.2.4, we have a contradiction. If instead $|T_{2n}| \geq n$, a similar argument shows that we derive a contradiction. Thus, such a $k$ exists and note that in particular $k$ must be odd.

Order the set of local minima $\{\delta_{j_{k-4n+1}}, \delta_{j_{k-4n+3}}, \ldots, \delta_{j_{k+4n-1}}\}$ in increasing order as $\gamma_1, \ldots, \gamma_{4n}$. Let

$$A' = \{\delta_{j_{k-4n+1}}, \delta_{j_{k-4n+3}}, \ldots, \delta_{j_{k-1}}\} \text{ and } B' = \{\delta_{j_{k+1}}, \delta_{j_{k+3}}, \ldots, \delta_{j_{k+4n-1}}\}.$$

Note that since $A', B'$ partition $\{\delta_{j_{k-4n+1}}, \delta_{j_{k-4n+3}}, \ldots, \delta_{j_{k+4n-1}}\}$, either $|A' \cap \{\gamma_1, \ldots, \gamma_{2n}\}| \geq n$ or $|B' \cap \{\gamma_1, \ldots, \gamma_{2n}\}| \geq n$. Without loss of generality, we assume that the former occurs since a symmetric argument would follow otherwise. Then, we also have that $|B' \cap \{\gamma_{2n+1}, \ldots, \gamma_{4n}\}| \geq n$. Set

$$A = A' \cap \{\gamma_1, \ldots, \gamma_{2n}\} \text{ and } B = B' \cap \{\gamma_{2n+1}, \ldots, \gamma_{4n}\}.$$

Let $a \in A$ and $b \in B$. By definition, $\delta_{j_a} < \delta_{j_b}$, and note that $b < k + 4n \Rightarrow b + 1 \leq k + 4n$, so

$$\delta(v_{j_a}, v_{j_a+1}, v_{j_b}, v_{j_b+1}, v_{j_{b+1}+1}) = (\delta_{j_a} < \delta_{j_k} > \delta_{j_b} < \delta_{j_{b+1}}),$$

where $\delta_{j_k} > \delta_{j_{b+1}}$ by definition. Since

$$\chi(v_{j_a}, v_{j_a+1}, v_{j_b}, v_{j_b+1}, v_{j_{b+1}+1}) = \text{blue},$$

we cannot have both $\phi(\delta_{j_a}, \delta_{j_{b+1}}) = \text{red}$ and $\phi(\delta_{j_a}, \delta_{j_b}) = \text{blue}$. Finally, restricting to any $n$ elements of $A, B$ and letting

$$C = \{\delta_{j_{b+1}} : \delta_{j_b} \in B\},$$

and defining $f : B \to C$ via $\delta_{j_b} \mapsto \delta_{j_{b+1}}$, we obtain 3 disjoint $n$-sets with precisely the structure avoided in the graph coloring $\phi$, a contradiction.

Thus, $\chi$ does not produce a blue $K^{(5)}_{128n^4}$ on $V$. □

## 2.3 Concluding remarks

We have determined the tower growth rate for $r_k(k+1, k-1; n)$. Thus, the only problem remaining for the Erdős-Hajnal hypergraph Ramsey conjecture, is to determine the tower growth rate for $r_k(k+1,k;n)$.

Let us remark that similar arguments show that $r_5(6,5;4n^2) > 2^{r_4(5,4;n)-1}$. To define such a coloring, let $N = r_4(5,4;n)-1$ and let $\varphi$ be a red/blue coloring of the 4-tuples of $\{0,\dots,N-1\}$ such that there are there are at most 3 red edges among every 5 vertices and there is no blue clique of size $n$. We then color the 5-tuples of $V = \{0,1,\dots,2^N-1\}$ so that $\chi$ produces at most 4 red edges among any 6 vertices and $\chi$ does not produce a blue clique of size $4n^2$. For vertices $v_1,\dots,v_5$ with $v_1 < v_2 < \cdots < v_5$, let $\delta_i = \delta(v_i, v_{i+1})$. We set $\chi(v_1,\dots,v_5) = \text{red}$ if:

1. We have that $\delta_1,\delta_2,\delta_3,\delta_4$ are monotone and $\varphi(\delta_1,\delta_2,\delta_3,\delta_4) = \text{red}$.

2. We have that $\delta_1 > \delta_2 < \delta_3 > \delta_4$ and $\delta_1 < \delta_3$.

Together with Lemma 2.2.1, showing that $r_4(5,4;n)$ grows double exponential in a power of $n$ would thus show that $r_k(k+1,k;n) = \text{twr}_{k-1}(n^{\Theta(1)})$.

### 2.3.1 Acknowledgements

Chapter 1 is a reprint, in full, of the material of the paper *A note on the Erdős–Hajnal hypergraph Ramsey problem*, which was published in *Proceedings of the American Mathematics Society*, volume 150(9), 2022, pp. 3675–3685. This paper was co-authored by the dissertation author together with Dhruv Mubayi and Andrew Suk.

# Chapter 3

# Isomorphisms between dense random graphs

We consider two variants of the induced subgraph isomorphism problem for two independent binomial random graphs with constant edge-probabilities $p_1, p_2$. In particular, (i) we prove a sharp threshold result for the appearance of $G_{n,p_1}$ as an induced subgraph of $G_{N,p_2}$, (ii) we show two-point concentration of the size of the maximum common induced subgraph of $G_{N,p_1}$ and $G_{N,p_2}$, and (iii) we show that the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$ has an unusual limiting distribution.

These results confirm simulation-based predictions of McCreesh, Prosser, Solnon and Trimble, and resolve several open problems of Chatterjee and Diaconis. The proofs are based on careful refinements of the first and second moment method, using extra twists to (a) take some non-standard behaviors into account, and (b) work around the large variance issues that prevent standard applications of these methods.

## 3.1 Introduction

Applied benchmark tests for the famous 'subgraph isomorphism problem' empirically discovered interesting phase transitions in random graphs. More concretely, these phase transitions were observed in two induced variants of the 'subgraph containment problem' widely-studied in random graph theory. In this paper we prove that the behavior of these two new random graph problems is surprisingly rich, with unexpected phenomena such as (a) that the form of the answer changes for constant edge-probabilities, (b) that the classical second moment method fails due to large variance, and (c) that an unusual limiting distribution arises.

To add more context, in many applications such as pattern recognition, computer vision, biochemistry and molecular science, it is a fundamental problem to determine whether an induced copy of a given graph $F$ (or a large part of a given graph $F$) is contained in another graph $G$; see [CH99, RW02, CFSV04, DSdlH$^+$11, ER11, GBB$^+$13, BGP$^+$13, MPST18]. In this paper we consider two probabilistic variants of this problem, where the two graphs $F$ and $G$ are both independent binomial random graphs with constant edge-probabilities $p_1, p_2 \in (0,1)$. Our main results are threefold:

- We prove a sharp threshold result for the appearance of $G_{n,p_1}$ as an induced subgraph of $G_{N,p_2}$, and discover that the sharpness differs between the cases $p_2 = 1/2$ and $p_2 \neq 1/2$; see Theorem 3.1.1.

- We show that the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$ has a Poisson limiting distribution for $p_2 = 1/2$, and a 'squashed' log-normal limiting distribution for $p_2 \neq 1/2$; see Theorem 3.1.2.

- We show two-point concentration of the maximum common induced subgraph of $G_{N,p_1}$ and $G_{N,p_2}$, and discover that the form of the maximum size changes as we vary $p_1, p_2 \in (0,1)$; see Theorem 3.1.3.

These results confirm simulation-based phase transition predictions of McCreesh, Prosser, Solnon and Trimble [MPT16, MPST18], and resolve several open problems of Chatterjee and Diaconis [CD23]. Our proofs are based on careful refinements of the first and second moment method, using several extra twists to (a) take the non-standard phenomena into account, and (b) work around the large variance issues that prevent standard applications of these moment based methods, using in particular pseudorandom properties and multi-round exposure arguments to tame the variance.

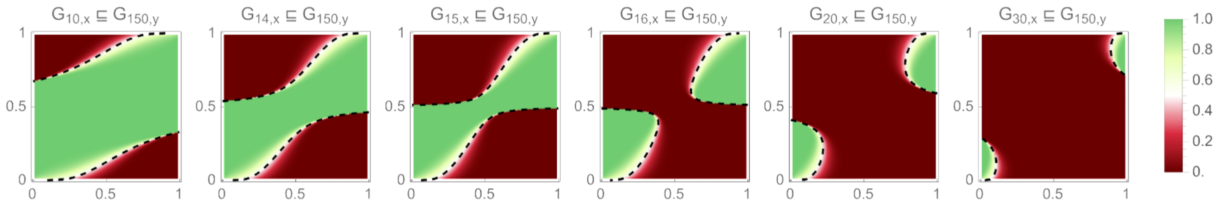



**Figure 3.1.** Theorem 3.1.1 establishes a sharp threshold around $n^* = n^*(p_1, p_2, N) := 2\log_a N + 1$ for the appearance of the binomial random graph $G_{n,p_1}$ as an induced subgraph of the independent random graph $G_{N,p_2}$. It also yields the induced containment probability estimate $\mathbb{P}(G_{n^*+c,p_1} \sqsubseteq G_{N,p_2}) \approx f_{p_1,p_2}(c)$, which allows us to reproduce an idealized version of Figure 5 in [MPST18], where $\mathbb{P}(G_{n,x} \sqsubseteq G_{N,y})$ with $N = 150$ is empirically plotted for all $x, y \in [0,1]$ and $n \in \{10, 14, 15, 16, 20, 30\}$; the dashed line corresponds to the threshold $n^*$. Previous work [CD23] applied to the special case $p_1 = p_2 = 1/2$, i.e., only reproduced the central point in each plot.

### 3.1.1 Induced subgraph isomorphism problem for random graphs

In the *induced subgraph isomorphism problem* the objective is to determine whether $F$ is isomorphic to an induced subgraph of $G$, i.e., whether $G$ contains an induced copy of $F$. In this paper we focus on this problem for two independent binomial random graphs $F = G_{n,p_1}$ and $G = G_{N,p_2}$, with constant edge-probabilities $p_1, p_2 \in (0,1)$ and $n \ll N$ many vertices. The motivation here is that, in applied work on benchmark tests for this NP-hard problem, it was empirically discovered [MPT16, MPST18] that such random graphs can be used to generate algorithmically hard problem-instances, leading to intriguing phase transitions. Knuth [Jan22, CD23] asked for mathematical explanations of these phase transitions, which are illustrated in Figure 3.1 via containment probability phase-diagram plots. The central points of these plots for $p_1 = p_2 = 1/2$ were resolved by Chatterjee and Diaconis [CD23] (and independently by Alon [Alo17, Alo23]), who emphasized in talks that the more interesting general case seems substantially more complicated. In this paper we resolve the general case with edge-probabilities $(p_1, p_2) \in (0,1)^2$: besides explaining the phase-diagram plots in Figure 3.1, we discover that the non-uniform case $p_2 \neq 1/2$ gives rise to new phenomena not anticipated by earlier work, including a different sharpness of the phase transition and an unusual 'squashed' log-normal limiting distribution.

Our first result establishes a sharp threshold for the appearance of the binomial random graph $G_{n,p_1}$ as an induced subgraph of the independent random graph $G_{N,p_2}$, resolving an open problem of Chatterjee and Diaconis [CD23]. Below the abbreviation $F \sqsubseteq G$ means that $G$ contains an induced copy of $F$.

**Theorem 3.1.1** (Sharp threshold)**.** *Let $p_1, p_2 \in (0,1)$ be constants. Define*

$$a := 1/\big(p_2^{p_1}(1-p_2)^{1-p_1}\big), \qquad\qquad \varepsilon_N := (\log\log N)^2/\log N.$$

*Then the following holds, for independent binomial random graphs $G_{n,p_1}$ and $G_{N,p_2}$:*

*(i) If $p_2 = 1/2$, then $a = 2$ and*

$$\lim_{N\to\infty} \mathbb{P}\big(G_{n,p_1} \sqsubseteq G_{N,p_2}\big) = \begin{cases} 1 & \text{if } n \le 2\log_a N + 1 - \varepsilon_N, \\ 0 & \text{if } n \ge 2\log_a N + 1 + \varepsilon_N. \end{cases} \tag{3.1}$$

*(ii) If $p_2 \neq 1/2$, then*

$$\lim_{N\to\infty} \mathbb{P}\big(G_{n,p_1} \sqsubseteq G_{N,p_2}\big) = \begin{cases} 1 & \text{if } n - (2\log_a N + 1) \to -\infty, \\ f(c) & \text{if } n - (2\log_a N + 1) \to c \in (-\infty, +\infty), \\ 0 & \text{if } n - (2\log_a N + 1) \to \infty, \end{cases} \tag{3.2}$$

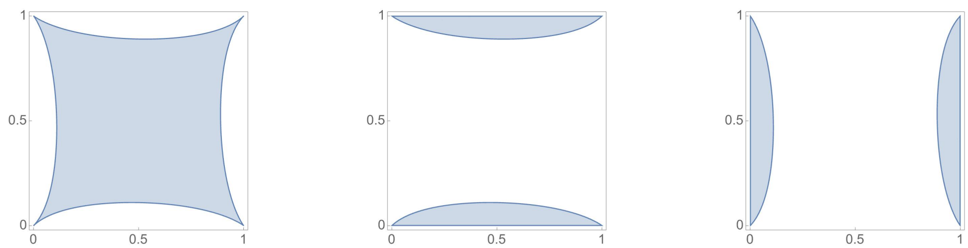
*where $f(c) = f_{p_1,p_2}(c) := \mathbb{P}\big(\mathrm{N}(0,\sigma^2) \ge c\big)$ for a normal random variable $\mathrm{N}(0,\sigma^2)$ with mean $0$ and variance $\sigma^2 := 2p_1(1-p_1)\log_a^2(1/p_2 - 1)$.*

In concrete words, Theorem 3.1.1 shows that, around the threshold $n \approx 2\log_a N + 1$, the induced containment probability $\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2})$ drops from $1 - o(1)$ to $o(1)$ in a window of size at most two when $p_2 = 1/2$, whereas this window has unbounded size when $p_2 \neq 1/2$. Our proof explains why this new phenomenon happens in the non-uniform case $p_2 \neq 1/2$: here the asymmetry of edges and non-edges makes $\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2} \mid G_{n,p_1})$ strongly dependent on the number of edges in $G_{n,p_1}$, which does not occur in the uniform case $p_2 = 1/2$ considered in previous work [CD23]. This edge-deviation effect turns out to be the driving force behind the limiting probability $f(c) \in (0,1)$ in (3.2); see Section 3.1.3 for more proof heuristics.

Theorem 3.1.1 confirms the simulation based predictions of McCreesh, Prosser, Solnon and Trimble [MPT16, MPST18], who empirically plotted the induced containment probability $\mathbb{P}(G_{n,x} \sqsubseteq G_{N,y})$ for $N = 150$ and predicted a phase transition near $n \approx 2\log_a N$; see Figure 5 and Section 3.1 in [MPST18]. Figure 3.1 illustrates that the fuzziness they found near their predicted threshold can be explained by the limiting probability $f(c)$ in (3.2), whose existence was not predicted in earlier work (of course, the 'small number of vertices' effect also leads to some discrepancies in the plots, in particular for very small and large edge-probabilities in Figure 3.1).

The classical problems of determining the size of the largest independent set and clique of $G_{N,p_2}$ are both related to Theorem 3.1.1, as they would correspond to the excluded edge-probabilities $p_1 \in \{0,1\}$. These two classical parameters $\alpha(G_{N,p_2})$ and $\omega(G_{N,p_2})$ are well-known [Bol01, JŁR00] to typically have size $2\log_a N - \Theta(\log\log N)$ for $a \in \{1/(1-p_2), 1/p_2\}$, and this additive $\Theta(\log\log N)$ left-shift compared to the threshold from Theorem 3.1.1 stems from an important conceptual difference: $n$-vertex cliques and independent sets have an automorphism group of size $n!$, whereas $G_{n,p_1}$ typically has a trivial automorphism group. This is one reason why our proof needs to take pseudorandom properties of random graphs into account; see also Sections 3.1.3 and 3.1.3.

As a consequence of our proof of Theorem 3.1.1, we are able to determine the asymptotic distribution of the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$, resolving another open problem of Chatterjee and Diaconis [CD23]. In concrete words, Theorem 3.1.2 (i) shows that the number of induced copies has a Poisson distribution for $p_2 = 1/2$ and $n$ close to the sharp threshold location $2\log_a N + 1$. Furthermore, Theorem 3.1.2 (ii) shows that the number of induced copies has a 'squashed' log-normal distribution for $p_2 \neq 1/2$ and $n = 2\log_a N + \Theta(1)$, which is a rather unusual limiting distribution for random discrete structures (that intuitively arises since the number of such induced copies is so strongly dependent on the number of edges in $G_{n,p_1}$; see




**(a)** **(b)** **(c)**

**Figure 3.2.** Theorem 3.1.3 establishes two-point concentration around $n_N$ of the maximum common induced subgraph of two independent binomial random graphs $G_{N,p_1}$ and $G_{N,p_2}$. The plots illustrate how the form of $n_N$ subdivides the unit square of edge-probabilities $(p_1,p_2)\in(0,1)^2$: we have $n_N = x_N^{(0)}(\hat{p}) \sim 4\log_{b_0(\hat{p})} N$ in region (a), $n_N \sim x_N^{(1)}(p_1^*) \sim 2\log_{b_1(p_1^*)} N$ in region (b), and $n_N \sim x_N^{(2)}(p_2^*) \sim 2\log_{b_2(p_2^*)} N$ in region (c), where $\hat{p}$ is defined as in (3.8) and $p_i^*$ is the unique solution of $\log b_0(p) = 2\log b_i(p)$; see Lemma 3.4.1 in Section 3.4. Previous work [CD23] applied to the special case $p_1 = p_2 = 1/2$, i.e., only to the central point in region (a).

Section 3.2.2).

**Theorem 3.1.2** (Asymptotic distribution)**.** *Let $p_1, p_2 \in (0,1)$ be constants. Define*

$$a := 1/\big(p_2^{p_1}(1-p_2)^{1-p_1}\big), \qquad\qquad \varepsilon_N := (\log\log N)^2/\log N$$

*as in Theorem 3.1.1. For independent binomial random graphs $G_{n,p_1}$ and $G_{N,p_2}$, let $X = X_{n,N}$ denote the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$. Then the following holds, as $N\to\infty$:*

*(i) If $p_2 = 1/2$ and $2\log_a N - 1 + \varepsilon_N \le n \le N$, then $X$ has asymptotically Poisson distribution with mean $\mu = \mu_{n,N} := (N)_n 2^{-\binom{n}{2}} = \big(N\cdot 2^{-(n-1)/2+O(n/N)}\big)^n$, i.e.,*

$$\mathrm{d_{TV}}\big(X,\, \mathrm{Po}(\mu)\big) \;\to\; 0. \tag{3.3}$$

*(ii) If $p_2 \neq 1/2$ and $n-(2\log_a N+1) \to c \in (-\infty,\infty)$ as $N\to\infty$, then*

$$\frac{\log(1+X)}{\log N} \overset{\mathrm{d}}{\to} \mathrm{SN}\Big(-c,\; 2p_1(1-p_1)\log_a^2(1/p_2-1)\Big), \tag{3.4}$$

*where $\mathrm{SN}(\mu,\sigma^2)$ has cumulative distribution function $F(x) := \mathbb{1}_{\{x\geq 0\}}\mathbb{P}(\mathrm{N}(\mu,\sigma^2)\le x)$.*

### 3.1.2 Maximum common induced subgraph problem for random graphs

In the *maximum common induced subgraph problem* the objective is to determine the maximum size of an induced subgraph of $F$ that is isomorphic to an induced subgraph of $G$, where the maximum size is with respect to the number of vertices (this generalizes the induced subgraph isomorphism problem, since the maximum common induced subgraph is $F$ if and only if $F$ is isomorphic to an induced subgraph of $G$). In this paper we focus on this problem for two independent binomial random graphs $F = G_{N,p_1}$ and $G = G_{N,p_2}$, with constant edge-probabilities $p_1, p_2 \in (0,1)$. One motivation here comes from combinatorial probability theory [CD23], where the following paradox was recently pointed out: two independent infinite random graphs with edge-probabilities $p_1, p_2 \in (0,1)$ are isomorphic with probability one, but two independent finite random graphs $G_{N,p_1}$ and $G_{N,p_2}$ are isomorphic with probability tending to zero as $N \to \infty$. This discontinuity of limits raises the question of finding the size of the maximum common induced subgraph of $G_{N,p_1}$ and $G_{N,p_2}$, which also is a natural random graph problem in its own right. Chatterjee and Diaconis [CD23] answered this question in the special case $p_1 = p_2 = 1/2$. In this paper we resolve the general case with edge-probabilities $(p_1, p_2) \in (0,1)^2$: we discover that the general form of the maximum size is significantly more complicated than for uniform random graphs, which in fact is closely linked to large variance difficulties.

Our next result establishes two-point concentration of the size $I_N$ of the maximum common induced subgraph of two independent binomial random graphs $G_{N,p_1}$ and $G_{N,p_2}$, resolving an open problem of Chatterjee and Diaconis [CD23]. In concrete words, Theorem 3.1.3 shows that $I_N$ equals, with probability tending to one as $N \to \infty$, one of (at most) two consecutive integers; see (3.5)–(3.6) below.

**Theorem 3.1.3** (Two-point concentration)**.** *Let $p_1, p_2 \in (0,1)$ be constants. For independent binomial random graphs $G_{N,p_1}$ and $G_{N,p_2}$, define $I_N$ as the size of the maximum common induced* *subgraph. Then*

$$\lim_{N\to\infty} \mathbb{P}\big(I_N \in \{\lfloor n_N - \varepsilon_N \rfloor, \lfloor n_N + \varepsilon_N \rfloor\}\big) = 1, \tag{3.5}$$

*where* $\varepsilon_N := (\log\log N)^2/\log N$ *and the parameter* $n_N = n_N(p_1, p_2)$ *defined in Remark 3.1.5 satisfies*

$$n_N = \frac{4\log N + O(\log\log N)}{\min_{p\in[0,1]} \max\big\{\log b_0(p),\, 2\log b_1(p),\, 2\log b_2(p)\big\}}, \tag{3.6}$$

*where, using the convention* $0^0 = 1$*, we have*

$$\begin{aligned} b_0 = b_0(p) &:= \left(\frac{p}{p_1 p_2}\right)^p \left(\frac{1-p}{(1-p_1)(1-p_2)}\right)^{1-p}, \\ b_i = b_i(p) &:= \left(\frac{p}{p_i}\right)^p \left(\frac{1-p}{1-p_i}\right)^{1-p}. \end{aligned} \tag{3.7}$$

Interestingly, Figure 3.2 shows that the form of the two-point concentration location $n_N = n_N(p_1, p_2)$ changes as we vary the edge-probabilities $(p_1, p_2) \in (0,1)^2$, which is a rather surprising phenomenon for random graphs with constant edge-probabilities. In Section 3.1.3 we discuss how the three different forms of $n_N$ heuristically arise (due to containment in $G_{N,p_1}$, containment in $G_{N,p_2}$, and containment in both). The value of $p$ which attains the minimum in (3.6) is an approximation of the edge-density of a maximum common induced subgraph in $G_{N,p_1}$ and $G_{N,p_2}$. There is a natural guess for the 'correct' edge-density: if we condition on $G_{n,p_1}$ and $G_{n,p_2}$ being equal, then using linearity of expectation the expected edge-density equals

$$\mathbb{E}\left(\frac{e(G_{n,p_1})}{\binom{n}{2}} \,\middle|\, G_{n,p_1} = G_{n,p_2}\right) = \frac{p_1 p_2}{p_1 p_2 + (1-p_1)(1-p_2)} =: \hat{p}(p_1, p_2) = \hat{p}. \tag{3.8}$$

It turns out that $\hat{p}(p_1, p_2)$ is indeed the correct edge-density for a range of edge-probabilities $(p_1, p_2) \in (0,1)^2$; see region (a) in Figure 3.2. In those cases Corollary 3.1.4 gives the two-point concentration location explicitly (the derivation is deferred to Section 3.4), which in the special case $p_1 = p_2 = 1/2$ recovers [CD23, Theorem 1.1].

**Corollary 3.1.4** (Special cases)**.** *Let* $p_1, p_2 \in (0,1)$ *be constants satisfying*

$$\log b_0(\hat{p}) > \max\big\{2\log b_1(\hat{p}),\, 2\log b_2(\hat{p})\big\}$$

*for $b_j(p)$ is as defined in* (3.7)*, which in particular holds when $p_1 = p_2$. For independent binomial random graphs $G_{N,p_1}$ and $G_{N,p_2}$, define $I_N$ as the size of the maximum common induced subgraph. Then*

$$\lim_{N\to\infty} \mathbb{P}\big(I_N \in \big\{ \lfloor x_N^{(0)}(\hat{p}) - \varepsilon_N \rfloor, \lfloor x_N^{(0)}(\hat{p}) + \varepsilon_N \rfloor \big\}\big) = 1, \tag{3.9}$$

*where $\varepsilon_N := (\log\log N)^2/\log N$, $\hat{p} = \hat{p}(p_1, p_2)$ is defined as in* (3.8)*, $b_0 = b_0(p)$ is defined as in* (3.7)*, and*

$$x_N^{(0)}(p) := 4\log_{b_0} N - 2\log_{b_0}\log_{b_0} N - 2\log_{b_0}(4/e) + 1. \tag{3.10}$$

In general, the two-point concentration location $n_N = n_N(p_1, p_2)$ is defined in (3.11) as the solution of an optimization problem over all edge-densities $p \in [0,1]$ of a potential maximum common subgraph. In Section 3.1.3 we discuss how this more complicated form of $n_N$ stems from large variance difficulties that can arise in the second moment method arguments (where the typical number of copies can be zero even when the expected number of copies tends to infinity). While the definition (3.11) of $n_N$ involves two implicitly defined (3.12) parameters $x_N^{(i)}(p)$, we remark that when $n_N = x_N^{(i)}(p)$ holds for $i \in \{1,2\}$, then $x_N^{(i)}(p)$ has the explicit form[1] given in (3.13) below; see Lemma 3.4.1 in Section 3.4 for further estimates of $n_N$.

*Remark* 3.1.5 (Explicit expression for $n_N$). The proof of Theorem 3.1.3 uses

$$n_N = n_N(p_1, p_2) := \max_{p\in[0,1]} \min\Big\{ x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p) \Big\}, \tag{3.11}$$

where $x_N^{(0)}(p)$ is as defined in (3.10), and $x_N^{(i)} = x_N^{(i)}(p)$ with $i \in \{1,2\}$ is the unique solution to

$$e \cdot N \;=\; x_N^{(i)}(p) \cdot b_i^{\big(x_N^{(i)}(p)-1\big)/2}, \tag{3.12}$$

where $b_i = b_i(p)$ is as defined in (3.7). Furthermore, if $n_N = x_N^{(i)}(p)$ holds in (3.11) for $i \in \{1,2\}$,

[1]One might be tempted to think that the estimate (3.13) could be used to explicitly define $x_N^{(i)}(p)$ for $i \in \{1,2\}$, by simply ignoring the additive error term. Unfortunately, this does not work for a subtle technical reason: for fixed $N$ this would lead to $\lim_{p\to p_i} x_N^{(i)}(p) = -\infty$ (since $\lim_{p\to p_i} b_i(p) = b_i(p_i) = 1$), making it inadequate for the definition (3.11) of $n_N$.

then

$$x_N^{(i)}(p) = 2\log_{b_i} N - 2\log_{b_i}\log_{b_i} N - 2\log_{b_i}(2/e) + 1 + O\left(\frac{\log\log N}{\log N}\right). \tag{3.13}$$

### 3.1.3 Intuition and proof heuristics

The proofs of our main results are based on involved applications of the first and second moment method, which each require an extra twist due to large variance (that prevents standard applications of these methods). In the following we highlight some of the key intuition and heuristics going into our arguments.

**Induced subgraph isomorphism problem**

We now heuristically discuss the sharp threshold for induced containment $G_{n,p_1} \sqsubseteq G_{N,p_2}$. Besides outlining the reasoning behind our proof approach for Theorem 3.1.1, we also motivate why the threshold is located around $n \approx 2\log_a N + 1$, and clarify why the case $p = 1/2$ behaves so differently than the case $p \neq 1/2$.

The natural proof approach would be to apply the first and second moment method to the random variable $X$ that counts the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$. While this approach can indeed be used to establish (3.1) when $p_2 = 1/2$ (as done in [CD23] for $p_1 = p_2 = 1/2$), it fails when $p_2 \neq 1/2$: the reason is $\mathrm{Var} X \gg (\mathbb{E}X)^2$, i.e., that the variance of $X$ is too large to apply the second moment method.

We overcome this second moment challenge for $p_2 \neq 1/2$ by identifying the key reason for the large variance, which turns out to be random fluctuations of the number of edges in $G_{n,p_1}$. To work around the effect of these edge-fluctuations we use a multi-round exposure approach, where we first reveal $G_{n,p_1}$ and then $G_{N,p_2}$ (which conveniently allows us to deal with one source of randomness at a time). When we reveal $G_{n,p_1}$, we exploit that $H := G_{n,p_1}$ will typically be asymmetric and satisfy other pseudorandom properties. Writing $X_H$ for the number of induced copies of $H$ in $G_{N,p_2}$, we then focus on the conditional probability

$$\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2} \mid G_{n,p_1} = H) = \mathbb{P}(H \sqsubseteq G_{N,p_2}) = \mathbb{P}(X_H \geq 1). \tag{3.14}$$

To see how the number of edges in $H$ affects this containment probability, suppose for concreteness that $H$ has $m = \binom{n}{2}p_1 + \delta n/2$ edges. Recalling $a = 1/\big(p_2^{p_1}(1-p_2)^{1-p_1}\big)$, it turns out (see (3.31)–(3.32) in Section 3.2.1 for the routine details) that the expected number of induced copies of $H$ in $G_{N,p_2}$ satisfies

$$\mathbb{E}X_H = (N)_n \cdot p_2^m(1-p_2)^{\binom{n}{2}-m} = \left[\left(\frac{p_2}{1-p_2}\right)^{\delta/2} N e^{-\log a\cdot(n-1)/2+O(n/N)}\right]^n. \tag{3.15}$$

For $n = 2\log_a N + 1 + c$ and $p_2 \neq 1/2$ it follows from (3.15) that the value of the edge-deviation parameter $\delta$ determines whether $\mathbb{E}X_H$ goes to $\infty$ or 0, i.e., depending on whether $\delta \log\big(p_2/(1-p_2)\big)$ is smaller or larger than $c\log a$ (here we also see why the case $p_2 = 1/2$ behaves so differently: in (3.15) the term involving $\delta$ equals one and thus disappears). With some technical effort, we can make the first moment heuristic that $\mathbb{E}X_H \to \infty$ implies $\mathbb{P}(X_H \geq 1) \to 1$ rigorous, i.e., use the first and second moment method to show that

$$\mathbb{P}(X_H \geq 1) = \begin{cases} o(1) & \text{if } \delta\log\big(p_2/(1-p_2)\big) < c\log a, \\ 1-o(1) & \text{if } \delta\log\big(p_2/(1-p_2)\big) > c\log a, \end{cases} \tag{3.16}$$

where a good control of the 'overlaps' of different $H$ in the variance calculations requires us to identify and exploit suitable pseudorandom properties of $H$ (which we can 'with foresight' insert into our argument when we first reveal $H = G_{n,p_1}$). The crux is that the containment conditions in (3.16) only depend on deviations in the number $e(G_{n,p_1}) = \binom{n}{2}p_1 + \delta n/2$ of edges in $G_{n,p_1}$, which in view of (3.14) intuitively translates into

$$\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2}) = \mathbb{P}\left(\frac{e(G_{n,p_1}) - \binom{n}{2}p_1}{n/2} \cdot \log\big(p_2/(1-p_2)\big) > c\log a\right) + o(1) \tag{3.17}$$

for $n = 2\log_a N + 1 + c$. This in turn makes the threshold result (3.2) plausible via the Central Limit Theorem (in fact, (3.17) is also consistent with the form of (3.1) for $p_2 = 1/2$, since then $\log\big(p_2/(1-p_2)\big) = 0$); see Section 3.2.1 for the full technical details of the proof of

Theorem 3.1.1.

With this knowledge about $X_H$ in hand, the distributional result Theorem 3.1.2 follows without much extra conceptual work. Indeed, for $p \neq 1/2$ we obtain the unusual limiting distribution (3.4) by exploiting that the event $X_H \geq 1$ in the second moment method based $1-o(1)$ statement in (3.16) can be strengthened to $X_H \approx \mathbb{E}X_H$; see Section 3.2.2. Furthermore, for $p = 1/2$ we obtain the Poisson limiting distribution (3.3) by refining our variance estimates in order to apply the Stein-Chen method to $X_H$; see Section 3.2.3.

**Maximum common induced subgraph problem**

We now heuristically discuss the two-point concentration of the size $I_N$ of the maximum common induced subgraph in $G_{N,p_1}$ and $G_{N,p_2}$. Our main focus is on motivating the 'correct' typical value $I_N \approx n_N$ in Theorem 3.1.3, which requires a refinement of the vanilla first and second moment approach. For ease of exposition, here we shall restrict our attention to the first-order asymptotics of $n_N = n_N(p_1, p_2)$ from (3.6).

Armed with the first and second moment insights for $X_H$ from the heuristic discussion in Section 3.1.3, the natural approach towards determining the typical value of $I_N$ would be to focus on the random variables $X_{n,m} := \sum_{H \in \mathscr{G}_{n,m}} X_H^{(1)} X_H^{(2)}$, where $\mathscr{G}_{n,m}$ denotes the set of all unlabeled graphs with $n$ vertices and $m$ edges, and $X_H^{(i)}$ counts the number of induced copies of $H$ in $G_{N,p_i}$. Here the crux is that

$$I_N = \max\big\{n : X_{n,m} \geq 1 \text{ for some } 0 \leq m \leq \tbinom{n}{2}\big\}. \tag{3.18}$$

Using $|\mathscr{G}_{n,m}| = \binom{\binom{n}{2}}{m} e^{O(n \log n)}$ and the independence of $G_{N,p_1}$ and $G_{N,p_2}$, similar to (3.15) it turns out (see (3.60) and (3.68) in Section 3.3.1 for the details) that for edge-density $p = p(m,n) := m/\binom{n}{2}$ we have

$$\mathbb{E}X_{n,m} = e^{O(n \log n)} \binom{\binom{n}{2}}{m} \prod_{i \in \{1,2\}} (N)_n p_i^m (1-p_i)^{\binom{n}{2}-m} = \Big[N^2 e^{-\frac{n}{2} \log b_0(p) + O(\log n)}\Big]^n. \tag{3.19}$$

Taking into account when these expectations go to 0 and $\infty$, in view of (3.18) and standard first moment heuristics it then is natural to guess[2] that the typical value of $I_N$ should be approximately

$$n_N \approx \max_{p\in[0,1]} 4\log_{b_0(p)} N = \frac{4\log N}{\min_{p\in[0,1]}\log b_0(p)} = 4\log_{b_0(\hat{p})} N, \tag{3.20}$$

where $\hat{p} = \hat{p}(p_1, p_2)$ from (3.8) turns out to be the unique minimizer of $\log b_0(p)$ from (3.7). For this choice of $n_N$ it turns out that we indeed typically have $I_N \approx n_N$ for a range of edge-probabilities $(p_1, p_2) \in (0,1)^2$, including the special case $p_1 = p_2$; see Corollary 3.1.4 and region (a) in Figure 3.2.

For general edge-probabilities $(p_1, p_2) \in (0,1)^2$ the 'correct' form of $n_N$ is more complicated than our first guess (3.20), and the key reason turns out to be that due to large variance we can have $\mathbb{P}(X_{n,m} \geq 1) \to 0$ despite $\mathbb{E}X_{n,m} \to \infty$. We overcome this difficulty by realizing that containment in $G_{N,p_i}$ can, from a 'first moment perspective', sometimes be harder than containment in both $G_{N,p_1}$ and $G_{N,p_2}$, i.e., that we can have $\mathbb{E}X^{(i)}_{n,m} \to 0$ for $X^{(i)}_{n,m} := \sum_{H\in\mathscr{G}_{n,m}} X^{(i)}_H$ despite $\mathbb{E}X_{n,m} \to \infty$. Similarly to (3.19) and (3.15) it turns out (see (3.58) in Section 3.3.1 for the details) that for edge-density $p = p(m,n) = m/\binom{n}{2}$ we have

$$\mathbb{E}X^{(i)}_{n,m} = e^{O(n\log n)}\binom{\binom{n}{2}}{m}\cdot (N)_n p_i^m (1-p_i)^{\binom{n}{2}-m} = \Big[Ne^{-\frac{n}{2}\log b_i(p) + O(\log n)}\Big]^n. \tag{3.21}$$

Note that $X_{n,m} \geq 1$ implies both $X^{(1)}_{n,m} \geq 1$ and $X^{(2)}_{n,m} \geq 1$. Taking into account when the two expectations in (3.21) go to 0, we thus refine our first guess (3.20) for the typical value of $I_N$ to approximately

$$\begin{aligned} n_N &\approx \max_{p\in[0,1]} \min\big\{4\log_{b_0(p)} N,\, 2\log_{b_1(p)} N,\, 2\log_{b_2(p)} N\big\} \\ &= \frac{4\log N}{\min_{p\in[0,1]}\max\big\{\log b_0(p),\, 2\log b_1(p),\, 2\log b_2(p)\big\}}. \end{aligned} \tag{3.22}$$

This turns out to be the 'correct' value of $n_N$ up to second order correction terms, see (3.6) and Figure 3.2. Indeed, with substantial technical effort (more involved than for the induced subgraph isomorphism problem) we can use refinements of the first and second moment method to prove two-point concentration of form $I_N \approx n_N$ for all edge-probabilities $(p_1, p_2) \in (0,1)^2$;

[2] The first moment heuristic for $n_N$ is as follows. For any $n \geq (1+\xi)n_N$, the definition of $n_N$ and (3.19) imply that $\sum_m \mathbb{E}X_{n,m} \leq n^2 \cdot N^{-\Theta(\xi n)} \to 0$. Similarly, for any $n \leq (1-\xi)n_n$ we have $\mathbb{E}X_{n,m} \geq N^{\Theta(\xi n)} \to \infty$ for $m = \binom{n}{2}\hat{p}$.

see Section 3.3 for the full technical details of the proof of Theorem 3.1.3.

**Pseudorandom properties of random graphs**

To enable the technical variance estimates in the proofs of our main results, we need to get good control over the 'overlaps' of different induced copies. To this end it will be important to take the pseudorandom properties of random graphs into account, by restricting our attention to graphs $H$ from the following three classes:

$\mathscr{A}_n$: The set of all $n$-vertex graphs $H$ where, for all vertex-subsets $L \subseteq [n]$ of size $|L| \geq n - n^{2/3}$, the induced subgraph $H[L]$ is asymmetric, i.e., satisfies $|\mathrm{Aut}(H[L])| = 1$.

$\mathscr{E}_{n,m}$: The set of all $n$-vertex graphs $H$ where, for all non-empty vertex-subsets $L \subseteq [n]$, the induced subgraph $H[L]$ contains $e(H[L]) = \binom{|L|}{2}m/\binom{n}{2} \pm n^{2/3}(n-|L|)$ edges.

$\mathscr{F}_{n,p}$: The set of all $n$-vertex graphs $H$ with $e(H) = \binom{n}{2}p \pm n^{4/3}$ edges.

In concrete words, the first property $\mathscr{A}_n$ formalizes the folklore fact that all very large subsets of random graphs are asymmetric; see [Bol01, ER63, KSV02]. The second and third properties $\mathscr{E}_{n,m}$ and $\mathscr{F}_{n,p}$ formalize the well-known fact that the edges of a random graph are well-distributed; see [BBSV19, Bol01, FKM14]. The following auxiliary lemma states that these are indeed typical properties of dense random graphs (we defer the proof to Section 3.5, since it is rather tangential to our main arguments). Below $G_{n,m}$ denotes the uniform random graph, which is chosen uniformly at random from all $n$-vertex graphs with exactly $m$ edges. Furthermore, the abbreviation whp (with high probability) means with probability tending to one as the number of vertices goes to infinity.

**Lemma 3.1.6** (Random graphs are pseudorandom)**.** *Let $p, \gamma \in (0,1)$ be constants. Then the following holds:*

*(i) For all $\gamma\binom{n}{2} \leq m \leq (1-\gamma)\binom{n}{2}$, the uniform random graph whp satisfies $G_{n,m} \in \mathscr{A}_n \cap \mathscr{E}_{n,m}$.*

*(ii) The binomial random graph whp satisfies $G_{n,p} \in \mathscr{A}_n \cap \mathscr{F}_{n,p}$.*

The reason for why we only consider the edge-property $\mathscr{E}_{n,m}$ for the uniform random graph $G_{n,m}$ is conceptual: the edges of $G_{n,m}$ are simply more well-distributed than the edges of the binomial random $G_{n,p}$ with $p := m/\binom{n}{2} \in [\gamma, 1-\gamma]$, where for large vertex-subsets $L \subseteq [n]$ of size $|L| = n - o(n^{1/3})$ we expect that typically $\left|e(G_{n,p}[L]) - \binom{|L|}{2}p\right| = \Omega\big(|L|\sqrt{p(1-p)}\big) \gg n^{2/3}(n-|L|)$ holds.

## 3.2 Induced subgraph isomorphism problem

This section is devoted to the induced subgraph isomorphism problem for two independent random graphs $G_{n,p_1}$ and $G_{N,p_2}$ with constant edge-probabilities $p_1, p_2 \in (0,1)$. It naturally splits into several parts: in the core Section 3.2.1 we prove the sharp threshold result Theorem 3.1.1, and in the subsequent Sections 3.2.2–3.2.3 we then separately prove the two parts of the distributional result Theorem 3.1.2.

### 3.2.1 Sharp Threshold: Proof of Theorem 3.1.1

In this section we prove Theorem 3.1.1, i.e., establish a sharp threshold for the appearance of an induced copy of $G_{n,p_1}$ in $G_{N,p_2}$, where $p_1, p_2 \in (0,1)$. Our proof-strategy is somewhat roundabout, since we need to deal with the difficulty that for $p_2 \neq 1/2$ the induced containment event $G_{n,p_1} \sqsubseteq G_{N,p_2}$ is sensitive towards small variations in the number of edges of $G_{n,p_1}$ (as discussed in Sections 3.1.1 and 3.1.3). For this reason we first prove a sharp threshold result for the event that $G_{N,p_2}$ contains an induced copy of a particular $n$-vertex pseudorandom graph $H$ with $m$ edges (see Lemma 3.2.1 below). Since $G_{n,p_1}$ will typically be pseudorandom by Lemma 3.1.6, this then effectively reduces the problem of estimating $\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2})$ to understanding deviations in the number of edges of $G_{n,p_1}$, which is a well-understood and much simpler problem.

Turning to the details, to restrict our attention to graphs with pseudorandom properties we introduce

$$\mathscr{T}_{n,m} := \mathscr{A}_n \cap \mathscr{E}_{n,m}, \tag{3.23}$$

where the asymmetry property $\mathscr{A}_n$ and the edge-distribution property $\mathscr{E}_{n,m}$ are defined as in Section 3.1.3. For edge-counts $m$ of interest, Lemma 3.1.6 shows that the uniform random graph whp satisfies $G_{n,m} \in \mathscr{T}_{n,m}$, so Lemma 3.2.1 effectively gives a sharp threshold result for the induced containment $G_{n,m} \sqsubseteq G_{N,p_2}$.

**Lemma 3.2.1** (Sharp threshold for pseudorandom graphs)**.** *Let $p_1, p_2 \in (0,1)$ be constants. Define $a := 1/(p_2^{p_1}(1-p_2)^{1-p_1})$, $\varepsilon_N := (\log\log N)^2/\log N$, and $\psi := \log_a\big(p_2/(1-p_2)\big)$. If $\delta_m := \big[m - \binom{n}{2}p_1\big]/(n/2)$ and $c_N := n - (2\log_a N + 1)$ satisfy $|\psi\delta_m| = o(n)$ and $|c_N| = o(\log_a N)$, then the following holds: as $N \to \infty$, for any graph $H \in \mathscr{T}_{n,m}$:*

$$\mathbb{P}\big(H \sqsubseteq G_{N,p_2}\big) = \begin{cases} o(1) & \text{if } \psi\delta_m - c_N \le -\varepsilon_N, \\ 1 - o(1) & \text{if } \psi\delta_m - c_N \ge \varepsilon_N. \end{cases} \tag{3.24}$$

*Remark* 3.2.2. The $o(1)$ statement in (3.24) remains valid for any $n$-vertex graph $H$ with $e(H) = m$ edges.

Before giving the (first and second moment based) proof of Lemma 3.2.1, we first show how it implies Theorem 3.1.1 using a multi-round exposure argument, where we first reveal the number of edges of $G_{n,p_1}$, then reveal the random graph $G_{n,p_1}$, and afterwards try to embed $G_{n,p_1}$ into the random graph $G_{N,p_2}$, each time discarding atypical outcomes along the way (so that we can focus on pseudorandom $G_{n,p_1}$).

*Proof of Theorem 3.1.1.* To estimate the probability that $G_{n,p_1} \sqsubseteq G_{N,p_2}$ holds, by monotonicity in $n$ it suffices to consider the case where $c_N = o(\log\log N)$ holds (in fact $c_N = O(1)$ suffices, since $\lim_{c\to\infty} f(c) = 0$ and $\lim_{c\to-\infty} f(c) = 1$ when $p_2 \neq 1/2$). Since $G_{n,p_1}$ conditioned on having $m$ edges has the same distribution as $G_{n,m}$, by using Lemma 3.1.6 to handle outcomes of $G_{n,p_1}$ with an atypical number of edges $e(G_{n,p_1})$ it follows that

$$\left|\mathbb{P}\big(G_{n,p_1} \sqsubseteq G_{N,p_2}\big) - \sum_{m:\left|m-\binom{n}{2}p_1\right|\le n^{4/3}} \mathbb{P}\big(G_{n,m} \sqsubseteq G_{N,p_2}\big)\mathbb{P}\big(e(G_{n,p_1}) = m\big)\right|$$
$$\le \mathbb{P}(G_{n,p_1} \notin \mathscr{F}_{n,p_1}) = o(1). \tag{3.25}$$

If $\left|m - \binom{n}{2}p_1\right| \le n^{4/3}$, then by using Lemma 3.1.6 to handle the atypical event $G_{n,m} \notin \mathscr{T}_{n,m}$ it follows that

$$\left|\mathbb{P}(G_{n,m} \sqsubseteq G_{N,p_2}) - \sum_{H\in\mathscr{T}_{n,m}} \mathbb{P}(H \sqsubseteq G_{N,p_2})\mathbb{P}(G_{n,m} = H)\right| \le \mathbb{P}(G_{n,m} \notin \mathscr{T}_{n,m}) = o(1). \tag{3.26}$$

In particular, using the sharp threshold result Lemma 3.2.1 for the event $H \sqsubseteq G_{N,p_2}$ (which applies since $|\psi\delta_m| = o(n)$ holds) and Lemma 3.1.6 for the typical event $G_{n,m} \in \mathscr{T}_{n,m}$, it follows that (3.26) implies

$$\mathbb{P}(G_{n,m} \sqsubseteq G_{N,p_2}) = \begin{cases} o(1) & \text{if } \psi\delta_m - c_N \le -\varepsilon_N, \\ 1-o(1) & \text{if } \psi\delta_m - c_N \ge \varepsilon_N. \end{cases} \tag{3.27}$$

If $p_2 = 1/2$, then in the setting of (3.1) we have $\psi = 0$ and $|c_N| \ge \varepsilon_N$, so the sharp threshold result (3.1) follows by first inserting estimate (3.27) into (3.25) and finally using Lemma 3.1.6 to infer $\mathbb{P}(G_{n,p_1} \in \mathscr{F}_{n,p_1}) = 1 - o(1)$.

We focus on the case $p_2 \neq 1/2$ in the remainder of the proof. Note that estimate (3.27) does not apply when $|\psi\delta_m - c_N| < \varepsilon_N$, but we shall now argue that we can effectively ignore this small range of edge-counts $m$ in (3.25) above. Recall that $n \to \infty$ as $N \to \infty$. By the Central Limit Theorem it thus follows that, as $N \to \infty$,

$$Z := \psi\delta_{e(G_{n,p_1})} = \psi \cdot \frac{e(G_{n,p_1}) - \binom{n}{2}p_1}{n/2} \xrightarrow{\text{d}} \mathrm{N}\big(0,\, 2p_1(1-p_1)\psi^2\big). \tag{3.28}$$

Recalling that $\varepsilon_N = o(1)$ and $0 < \operatorname{Var} Z = 2p_1(1-p_1)\psi^2 = \Theta(1)$, it follows that

$$\mathbb{P}(|Z - c_N| < \varepsilon_N) \,=\, o(1). \tag{3.29}$$

(This alternatively follows from $\max_k \mathbb{P}(e(G_{n,p_1}) = k) \le O(1)/\sqrt{n^2p_1(1-p_1)} \le O(1/n)$, say.)

Inserting estimates (3.27) and (3.29) into (3.25), using Lemma 3.1.6 to infer $\mathbb{P}(G_{n,p_1} \in \mathscr{F}_{n,p_1}) = 1 - o(1)$ it follows that

$$\left|\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2}) - \mathbb{P}(Z - c_N \geq 0)\right| = o(1), \tag{3.30}$$

which together with the convergence result (3.28) and $\varepsilon_N = o(1)$ establishes the threshold result (3.2). □

In the following proof of Lemma 3.2.1 we shall apply the first and second moment method to the random variable $X_H$, which we define as the number of induced copies of $H$ in $G_{N,p_2}$. Here the restriction to pseudorandom graphs $H \in \mathscr{T}_{n,m}$ will be key for controlling the expectation $\mathbb{E}X_H$ and variance $\mathrm{Var}X_H$.

*Proof of Lemma 3.2.1.* Let $H$ be any $n$-vertex graph with $e(H) = m$ edges. Note that $X_H = \sum_{S \in \binom{[N]}{n}} I_S$, where $I_S$ is the indicator random variable for the event that the induced subgraph $G_{N,p_2}[S]$ is isomorphic to $H$. Note that there are exactly $n!/|\mathrm{Aut}(H)|$ distinct embeddings of $H$ into $S$. Using linearity of expectation and the standard shorthand $(N)_n := N!/(N-n)!$, in view of $m = \binom{n}{2}p_1 + \delta_m n/2$, $a = 1/(p_2^{p_1}(1-p_2)^{1-p_1})$, $p_2/(1-p_2) = a^{\psi}$ and $(n-1)/2 = \log_a N + c_N/2$ it follows that the expected number of induced copies of $H$ in $G_{N,p_2}$ satisfies

$$\begin{aligned}
\mathbb{E}X_H = \sum_{S \in \binom{[N]}{n}} \mathbb{E}I_S &= \binom{N}{n} \cdot \frac{n!}{|\mathrm{Aut}(H)|} \cdot p_2^{e(H)}(1-p_2)^{\binom{n}{2}-e(H)} \\
&= \frac{1}{|\mathrm{Aut}(H)|} \cdot (N)_n \cdot p_2^m (1-p_2)^{\binom{n}{2}-m} \\
&= \frac{1}{|\mathrm{Aut}(H)|} \cdot \left[ N e^{O(n/N)} \cdot \left(\frac{p_2}{1-p_2}\right)^{\delta_m/2} a^{-(n-1)/2} \right]^n \\
&= \frac{1}{|\mathrm{Aut}(H)|} \cdot \left[ a^{\psi\delta_m - c_N + o(\varepsilon_N)} \right]^{n/2},
\end{aligned} \tag{3.31}$$

where we used $n/N = \Theta((\log N)/N) = o(\varepsilon_N)$ and $1 < a = O(1)$ for the last step. If we have $\psi\delta_m - c_N \leq -\varepsilon_N$, then using $|\mathrm{Aut}(H)| \geq 1$ and Markov's inequality together with $\varepsilon_N n \to \infty$ as $N \to \infty$, we see that

$$\mathbb{P}(X_H > 0) \leq \mathbb{E}X_H \leq a^{-\Omega(\varepsilon_N n)} = o(1),$$

establishing the $o(1)$ statement in (3.24).

In the remainder we focus on the $1-o(1)$ statement in (3.24), i.e., we henceforth fix $H \in \mathscr{T}_{n,m}$ and assume that $\psi\delta_m - c_N \geq \varepsilon_N$. Since $H \in \mathscr{T}_{n,m} \subseteq \mathscr{A}_N$ implies $|\mathrm{Aut}(H)| = 1$, using estimate (3.31) we infer that

$$\mu := \mathbb{E}X_H = (N)_n \cdot p_2^m (1-p_2)^{\binom{n}{2}-m} = \Big[a^{\psi\delta_m - c_N + o(\varepsilon_N)}\Big]^{n/2} \geq a^{\Omega(\varepsilon_N n)} \to \infty. \tag{3.32}$$

To complete the proof of (3.24), using Chebyshev's inequality it thus suffices to show that $\mathrm{Var}\, X_H = o\big((\mathbb{E}X_H)^2\big)$. Recall that $X_H = \sum_{S \in \binom{[N]}{n}} I_S$, where $I_S$ is the indicator random variable for the event that the induced subgraph $G_{N,p_2}[S]$ is isomorphic to $H$. Since $I_R$ and $I_S$ are independent when $|R \cap S| \leq 1$, we have

$$\mathrm{Var}\, X_H \;\leq\; \sum_{2 \leq \ell \leq n} \underbrace{\sum_{R,S \in \binom{[N]}{n}: |R\cap S| = \ell} \mathbb{E}(I_R I_S)}_{=: w_\ell} \;=\; \sum_{2 \leq \ell < n} w_\ell + \underbrace{\mathbb{E}X_H}_{=\mu}. \tag{3.33}$$

To bound the parameter $w_\ell$ defined in (3.33) for $2 \leq \ell < n$, we first note that there are $\binom{N}{n}\binom{n}{\ell}\binom{N-n}{n-\ell}$ ways of choosing $R, S \in \binom{[N]}{n}$ with $|R \cap S| = \ell$. To then get a handle on $\mathbb{E}(I_R I_S)$, we count the number of ways we can embed two induced copies of $H$ into $R \cup S$ so that $G_{N,p_2}[R]$ and $G_{N,p_2}[S]$ are both isomorphic to $H$: there are $n!$ ways of embedding an induced copy of $H$ into $R$, at most $\binom{n}{n-\ell} \cdot (n-\ell)! = (n)_{n-\ell}$ ways of choosing and embedding $n-\ell$ vertices of a second induced copy of $H$ into $S \setminus R$, and at most $\max_{H \in \mathscr{T}_{n,m}, |L| = \ell} |\mathrm{Aut}(H[L])|$ ways of embedding the remaining $\ell$ vertices of the second induced copy of $H$ into $R \cap S$, where the maximum is taken over all vertex-subsets $L \subseteq V(H)$ of size $|L| = \ell$. As these embeddings determine all edges and non-edges in $G_{N,p_2}[R \cup S]$, it follows that

$$\begin{aligned} w_\ell \leq \binom{N}{n}\binom{n}{\ell}\binom{N-n}{n-\ell} \cdot n! \cdot (n)_{n-\ell} \cdot \max_{H \in \mathscr{T}_{n,m},\, |L| = \ell} |\mathrm{Aut}(H[L])| \\ \cdot \max_{H \in \mathscr{T}_{n,m},\, |L| = \ell} p_2^{2m - e(H[L])} (1-p_2)^{2\binom{n}{2} - \binom{\ell}{2} - 2m + e(H[L])}, \end{aligned} \tag{3.34}$$

where the two maxima in (3.34) are each taken over all vertex-subsets $L \subseteq V(H)$ of size $|L| = \ell$, as above. Note that $n = (2+o(1))\log_a N$, which implies $N - n > n - \ell$ for all sufficiently large $N$.

Recalling that $\mu = (N)_n p_2^m (1-p_2)^{\binom{n}{2}-m}$ by (3.32), using $n^2/N = o(1)$ it follows that

$$\frac{w_\ell}{\mu^2} \le \underbrace{\frac{(N-n)_{n-\ell}}{(N)_n}}_{=(1+o(1))N^{-\ell}} \cdot \binom{n}{\ell}^2 \cdot \max_{H\in\mathscr{T}_{n,m},\,|L|=\ell} |\mathrm{Aut}(H[L])| \cdot \underbrace{\max_{H\in\mathscr{T}_{n,m},\,|L|=\ell} \left(\frac{1-p_2}{p_2}\right)^{e(H[L])} (1-p_2)^{-\binom{\ell}{2}}}_{=:P_\ell}. \tag{3.35}$$

Setting $\zeta := 1/\max\{-2\log_a(1-p_2), -2\log_a(p_2), 2\} > 0$, we now bound $w_\ell/\mu^2$ further using a case distinction.

**Case 1.** $2 \le \ell \le \zeta n$.

Here we use $0 \le e(H[L]) \le \binom{\ell}{2}$ to deduce that the parameter $P_\ell$ from (3.35) satisfies

$$P_\ell \le \max_{0\le k\le \binom{\ell}{2}} p_2^{-k}(1-p_2)^{-\binom{\ell}{2}+k} \le \left[\max\left\{\frac{1}{p_2}, \frac{1}{1-p_2}\right\}\right]^{\binom{\ell}{2}}. \tag{3.36}$$

Inserting this estimate and the trivial bound $|\mathrm{Aut}(H[L])| \le |L|!$ into (3.35), using $\binom{n}{\ell} \le n^\ell/\ell!$ together with $n = (2+o(1))\log_a N$ and $\ell \le \zeta n \le (1+o(1))\log_{\max\{1/p_2, 1/(1-p_2)\}} N$, it follows for all large enough $N$ that

$$\frac{w_\ell}{\mu^2} \le 2N^{-\ell} \cdot \binom{n}{\ell}^2 \ell! \cdot P_\ell \le \left[N^{-1} n^2 \cdot \left[\max\left\{\frac{1}{p_2}, \frac{1}{1-p_2}\right\}\right]^{(\ell-1)/2}\right]^\ell \le N^{-\ell/3} \le a^{-\Omega(n)}. \tag{3.37}$$

**Case 2.** $\zeta n \le \ell < n$.

Here we improve the estimate (3.36) for $P_\ell$ using that $H \in \mathscr{T}_{n,m} \subseteq \mathscr{E}_{n,m}$: this pseudorandom edge-property implies that, for any vertex-subset $L \subseteq [n]$ of size $|L| = \ell$, the induced number of edges satisfies $e(H[L]) = \binom{\ell}{2} m / \binom{n}{2} + O(n^{2/3}(n-\ell))$. Recalling $a = 1/(p_2^{p_1}(1-p_2)^{1-p_1})$, using $m/\binom{n}{2} = p_1 + \delta_m/n + O(1/n^2)$ and $(1-p_2)/p_2 = a^{-\psi}$ it follows that the parameter $P_\ell$ from (3.35) satisfies

$$P_\ell = a^{\binom{\ell}{2}} \cdot \left(\frac{1-p_2}{p_2}\right)^{\binom{\ell}{2}\delta_m/n + O(n^{2/3}(n-\ell))} = a^{(1-\psi\delta_m/n)\binom{\ell}{2} + O(\psi n^{2/3}(n-\ell))}. \tag{3.38}$$

For $\zeta n \le \ell < n - n^{2/3}$ we then estimate $|\mathrm{Aut}(H[L])| \le \ell! \le n^\ell \le e^{O(n^{2/3}(n-\ell))}$, and for $n - n^{2/3} \le \ell < n$ we use $H \in \mathscr{T}_{n,m} \subseteq \mathscr{A}_n$ to obtain $\max_{|L|=\ell} |\mathrm{Aut}(H[L])| = 1$. Note that $1 < a = O(1)$

and $\psi = O(1)$. We now insert these estimates and $\binom{n}{\ell} = \binom{n}{n-\ell} \le n^{n-\ell}$ into (3.35). Then, using $\ell - 1 = (n-1) - (n-\ell)$ and $|\psi\delta_m/n| = o(1)$ as well as $(n-1)/2 = \log_a N + c_N/2$ and $n - \ell \ge 1$, it follows for all large enough $N$ that

$$
\begin{aligned}
\frac{w_\ell}{\mu^2} &\le 2N^{-\ell} n^{2(n-\ell)} \cdot a^{(1-\psi\delta_m/n)\binom{\ell}{2}} \cdot e^{O(n^{2/3}(n-\ell))} \\
&\le \Big[N^{-1} a^{(1-\psi\delta_m/n)(\ell-1)/2} \cdot e^{O(n^{2/3}(n-\ell)/\ell)}\Big]^{\ell} \\
&\le \Big[N^{-1} a^{(n-1)/2-\psi\delta_m/2+o(1)-(1-o(1))(n-\ell)/2} \cdot e^{o(n-\ell)}\Big]^{\ell} \\
&\le \Big[a^{c_N-\psi\delta_m}\Big]^{\ell/2} \le a^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)},
\end{aligned}
\tag{3.39}
$$

where for the second last inequality we used that $c_N - \psi\delta_m = -(\psi\delta_m - c_N) \le -\varepsilon_N$.

To sum up: inserting the estimates (3.37) and (3.39) into the variance bound (3.33), using (3.32) it follows that

$$
\frac{\operatorname{Var} X_H}{(\mathbb{E}X_H)^2} \le \frac{\sum_{2\le\ell<n} w_\ell + \mu}{\mu^2} \le o(1) + \frac{1}{\mu} = o(1),
$$

which completes the proof of Lemma 3.2.1, as discussed. □

*Remark* 3.2.3. If $H \in \mathscr{T}_{n,m}$ and $\psi\delta_m - c_N \ge \varepsilon_N$, then the above proof of Lemma 3.2.1 shows that

$$
\mathbb{P}\big(|X_H - \mathbb{E}X_H| < \xi\mathbb{E}X_H\big) = 1 - o(1)
$$

for any $\xi > 0$, where the expected value $\mathbb{E}X_H$ satisfies the asymptotic estimate (3.32).

### 3.2.2 Asymptotic Distribution: Proof of Theorem 3.1.2 (ii)

In this section we prove the distributional result Theorem 3.1.2 (ii) as a corollary of the results from Section 3.2.1, i.e., establish that for $p_2 \ne 1/2$ the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$ has a 'squashed' log-normal limiting distribution. Here the crux is $X_H$ is strongly dependent on the number of edges: indeed, $X_H$ quickly changes from 0 to $(1+o(1))\mathbb{E}X_H \to \infty$ as the number of edges $m = e(H)$ passes through the threshold $\psi\delta_m \sim c_N$, see Remarks 3.2.2 and 3.2.3. This makes it plausible that $\log(1+X_H)/\log N$ changes abruptly around that threshold, which together with $c_N \to c$ and the normal convergence result (3.28) for $\psi\delta_{e(G_{n,p_1})}$ intuitively explains the form of the limiting distribution of $\log(1+X_{G_{n,p_1}})/\log N$ in the convergence result (3.4).

*Proof of Theorem 3.1.2 (ii).* Note that $c_N \to c$ and thus $c_N = O(1)$ by assumption. In view of Remark 3.2.3, define $\mathscr{T}_n$ as the union of all $\mathscr{T}_{n,m}$ from Section 3.2.1 with $\left|m - \binom{n}{2}p_1\right| \le n^{4/3}$. With analogous reasoning as for (3.25)–(3.26), using Lemma 3.1.6 it follows that

$$\mathbb{P}(G_{n,p_1} \notin \mathscr{T}_n) \le \mathbb{P}(G_{n,p_1} \notin \mathscr{F}_{n,p_1}) + \max_{m: \left|m - \binom{n}{2}p_1\right| \le n^{4/3}} \mathbb{P}(G_{n,m} \notin \mathscr{T}_{n,m}) = o(1). \tag{3.40}$$

Furthermore, with an eye on the form (3.32) of $\mathbb{E}X_H$ in Lemma 3.2.1, as in (3.28) we have

$$Z := \psi \cdot \delta_{e(G_{n,p_1})} = \psi \cdot \frac{e(G_{n,p_1}) - \binom{n}{2}p_1}{n/2} \stackrel{\mathrm{d}}{\to} \mathrm{N}\big(0, 2p_1(1-p_1)\psi^2\big). \tag{3.41}$$

We now condition on $G_{n,p_1} = H \in \mathscr{T}_n$, so that $X|_{G_{n,p_1}=H} = X_H$ and $|Z| \le 2|\psi|n^{1/3} = o(\sqrt{\log N})$. In particular, if $Z - c_N \le -\varepsilon_N$ holds, then by applying Lemma 3.2.1 it follows that whp

$$\frac{\log(1+X_H)}{\log N} = \frac{\log 1}{\log N} = 0. \tag{3.42}$$

Furthermore, if $Z - c_N \ge \varepsilon_N$ holds, then by combining Remark 3.2.3 and estimate (3.32) with $c_N = O(1)$ and $|Z| = o(\sqrt{\log N})$ as well as $n = 2\log_a N + O(1)$ and $a = O(1)$ it follows that whp

$$\frac{\log(1+X_H)}{\log N} = \frac{\log\big(\mathbb{E}X_H\big)}{\log N} + \frac{O(1)}{\log N} = \frac{n(Z-c_n)}{2\log_a N} + \frac{O(1)}{\log N} = Z - c_N + \frac{o(\sqrt{\log N})}{\log N}. \tag{3.43}$$

Finally, by combining estimates (3.40) and (3.29) with the conclusions of (3.42)–(3.43), it follows that

$$\mathbb{P}\left(\left|\frac{\log(1+X)}{\log N} - \big(Z - c_n\big)\mathbb{1}_{\{Z - c_N \ge \varepsilon_N\}}\right| \ge \frac{1}{\sqrt{\log N}}\right) = o(1),$$

which together with (3.41) as well as $c_N \to c$ and $\varepsilon_N \to 0$ establishes the convergence result (3.4).

□

Note that the above proof only uses the first and second moment method, i.e., does not require the asymptotics of $\mathrm{Var} X_H$. Given the somewhat complicated limiting distribution of $X_H$, we leave it as an interesting open problem to complement (3.4) with near-optimal estimates on the rate of convergence.

### 3.2.3 Asymptotic Poisson Distribution: Proof of Theorem 3.1.2 (i)

In this section we complete the proof of Theorem 3.1.2 by proving Theorem 3.1.2 (i), i.e., establishing that for $p_2 = 1/2$ the number of induced copies of $G_{n,p_1}$ in $G_{N,p_2}$ has asymptotically Poisson distribution. To this end we shall use a version of the Stein-Chen method for Poisson approximation together with a two-round exposure argument and a refinement of the variance estimates from Section 3.2.1 for $p_2 = 1/2$.

*Proof of Theorem 3.1.2 (i).* Note that $X = X_{n,N}$ conditional on $G_{n,p_1} = H$ has the same distribution as $X_H$. This enables us to again use a two-round exposure argument, where we first reveal $G_{n,p_1}$ and then afterwards count the number of induced copies of $G_{n,p_1}$ in $G_{N,1/2}$. To this end, let $\mathscr{G}_n$ be the set of all $n$-vertex graphs. Together with $\mathscr{A}_n \subseteq \mathscr{G}_n$ as defined in Section 3.1.3, by applying Lemma 3.1.6 it follows that

$$\begin{aligned}
d_{\mathrm{TV}}\big(X, \mathrm{Po}(\mu)\big) &= \sup_{S\subseteq\mathbb{N}} \big|\mathbb{P}(X\in S) - \mathbb{P}(\mathrm{Po}(\mu)\in S)\big| \\
&= \sup_{S\subseteq\mathbb{N}} \Big|\sum_{H\in\mathscr{G}_n} \mathbb{P}(G_{n,p_1} = H)\cdot\big[\mathbb{P}(X_H\in S) - \mathbb{P}(\mathrm{Po}(\mu)\in S)\big]\Big| \\
&\le \sum_{H\in\mathscr{A}_n} \mathbb{P}(G_{n,p_1} = H)\cdot \sup_{S\subseteq\mathbb{N}}\big|\mathbb{P}(X_H\in S) - \mathbb{P}(\mathrm{Po}(\mu)\in S)\big| + \mathbb{P}(G_{n,p_1}\notin\mathscr{A}_n) \\
&\le \max_{H\in\mathscr{A}_n} d_{\mathrm{TV}}\big(X_H, \mathrm{Po}(\mu)\big) + o(1).
\end{aligned}$$

It thus remains to show that $d_{\mathrm{TV}}(X_H, \mathrm{Po}(\mu)) = o(1)$ for any $H \in \mathscr{A}_n$. Fix a graph $H \in \mathscr{A}_n$. As in Section 3.2.1 we write $X_H = \sum_{S\in\binom{[N]}{n}} I_S$, where $I_S$ is the indicator random variable for the event that $G_{N,p_2}[S]$ is isomorphic to $H$. Since $H \in \mathscr{A}_n$ implies $|\mathrm{Aut}(H)| = 1$, using (3.31) with $p_2 = 1/2$ it follows that

$$\mathbb{E}X_H = (N)_n \left(\tfrac{1}{2}\right)^{\binom{n}{2}} = \mu \qquad \text{for all } H \in \mathscr{A}_n. \tag{3.44}$$

Note that $d_{\mathrm{TV}}(X_H, \mathrm{Po}(\mu)) = o(1)$ immediately follows when $\mu \to 0$ (since then $X_H$ and $\mathrm{Po}(\mu)$ are both whp zero). We may thus henceforth assume that $\mu = \Omega(1)$, which in view of $\mu = \big(N\cdot 2^{-(n-1)/2+O(n/N)}\big)^n$ and $n \ge 2\log_2 N - 1 + \varepsilon_N$ implies $n = 2\log_2 N + O(1)$. Since $I_R$ and $I_S$ are independent when $|R\cap S|\le 1$, by applying the well-known version of the Stein-Chen method for Poisson approximation (based on so-called dependency graphs) stated in [JŁR00, Theorem 6.23] it routinely follows that

$$\mathrm{d_{TV}}\big(X_H,\mathrm{Po}(\mu)\big)\ \le\ \underbrace{\min\{\mu^{-1},1\}}_{=O(\mu^{-1})}\cdot\Bigg[\underbrace{\sum_{\substack{R,S\in\binom{[N]}{n}:\\ 2\le|R\cap S|<n}}\mathbb{E}(I_RI_S)}_{=:\Lambda_1}\ +\ \underbrace{\sum_{\substack{R,S\in\binom{[N]}{n}:\\ 2\le|R\cap S|\le n}}\mathbb{E}I_R\mathbb{E}I_S}_{=:\Lambda_2}\Bigg]. \tag{3.45}$$

To establish the convergence result (3.3), it thus remains to show that $\Lambda_1$ and $\Lambda_2$ are both $o(\mu)$.

We first estimate $\Lambda_2$ using basic counting arguments. In particular, note that

$$\Lambda_2=\sum_{R\in\binom{[N]}{n}}\mathbb{E}I_R\sum_{\substack{S\in\binom{[N]}{n}:\\ 2\le|R\cap S|\le n}}\mathbb{E}I_S\ \le\ \mu\cdot\sum_{2\le k\le n}\binom{n}{k}\binom{N-n}{n-k}n!2^{-\binom{n}{2}}.$$

Recalling that $n=2\log_2 N+O(1)$, for all $2\le k\le n-1$ we see (with room to spare) that

$$\frac{\binom{n}{k}\binom{N-n}{n-k}}{\binom{n}{k+1}\binom{N-n}{n-k-1}}=\frac{(k+1)(N-2n+k+1)}{(n-k)^2}\ge\frac{N-2n}{n}>1$$

for all sufficiently large $N$. Using $n=2\log_2 N+O(1)$ and $n\ge 2\log_2 N-1+\varepsilon_N$, it follows that

$$\begin{aligned}\frac{\Lambda_2}{\mu}&\le n\cdot\binom{n}{2}\binom{N-n}{n-2}\cdot n!2^{-\binom{n}{2}}\\ &\le n^5N^{n-2}\cdot 2^{-\binom{n}{2}}\\ &\le N^{-2}\cdot\Big(n^{5/n}N\cdot 2^{-(n-1)/2}\Big)^n\\ &\le O(2^{-n})\cdot\Big(2^{1-\varepsilon_N/2+o(\varepsilon_N)}\Big)^n\le 2^{-\Omega(\varepsilon_N n)}\le n^{-\Omega(\log n)}.\end{aligned}$$

Finally, we estimate $\Lambda_1$ from (3.45) by refining the variance estimates from the proof of Lemma 3.2.1. Namely, bounding the following parameter $w_\ell$ from (3.33) as in (3.34), by setting $p_2=1/2$ we infer that

$$w_\ell:=\sum_{\substack{R,S\in\binom{[N]}{n}:\\ |R\cap S|=\ell}}\mathbb{E}(I_RI_S)\le\binom{N}{n}\binom{n}{\ell}\binom{N-n}{n-\ell}\cdot n!\cdot(n)_{n-\ell}\cdot\max_{H\in\mathscr{T}_n,\,|L|=\ell}|\mathrm{Aut}(H[L])|\cdot 2^{\binom{\ell}{2}-2\binom{n}{2}}.$$

Using similar (but simpler) arguments as for inequality (3.35), in view of (3.44) it then follows that

$$\frac{w_\ell}{\mu} \le N^{n-\ell} \cdot \binom{n}{\ell}^2 \cdot \max_{H \in \mathscr{T}_n, |L|=\ell} |\mathrm{Aut}(H[L])| \cdot 2^{\binom{\ell}{2}-\binom{n}{2}}. \tag{3.46}$$

We now bound $w_\ell/\mu$ further using a case distinction.

**Case 1.** $2 \le \ell \le n - n^{2/3}$.

Here we insert the trivial bound $|\mathrm{Aut}(H[L])| \le \ell!$ into inequality (3.46). Writing

$$\tbinom{\ell}{2} - \tbinom{n}{2} = -(n-\ell)(n+\ell-1)/2,$$

using $\binom{n}{\ell} \le n^\ell/\ell!$ and $n \ge 2\log_2 N - 1 + \varepsilon_N$ it follows that

$$\begin{aligned}
\frac{w_\ell}{\mu} &\le N^{n-\ell} \cdot \binom{n}{\ell}^2 \cdot \ell! \cdot 2^{\binom{\ell}{2}-\binom{n}{2}} \\
&\le \Big(N \cdot n^{2\ell/(n-\ell)} \cdot 2^{-(n-1+\ell)/2}\Big)^{n-\ell} \\
&\le \Big(2^{2-\varepsilon_N-\ell(1-4(\log n)/(n-\ell))}\Big)^{(n-\ell)/2} \le 2^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)},
\end{aligned} \tag{3.47}$$

where for the second last inequality we optimized over all $2 \le \ell \le n - n^{2/3}$, using $(\log n)/n = o(\varepsilon_N)$.

**Case 2.** $n - n^{2/3} \le \ell < n$.

Here we exploit $H \in \mathscr{A}_n$ to insert $|\mathrm{Aut}(H[L])| = 1$ into inequality (3.46). Writing again $\binom{\ell}{2} - \binom{n}{2} = -(n-\ell)(n+\ell-1)/2$, using $\binom{n}{\ell} = \binom{n}{n-\ell} \le n^{n-\ell}$ and $\ell - 2 + \varepsilon_N = \Omega(n)$ it follows that

$$\begin{aligned}
\frac{w_\ell}{\mu} &\le N^{n-\ell} \cdot \binom{n}{\ell}^2 \cdot 2^{\binom{\ell}{2}-\binom{n}{2}} \\
&\le \Big(N \cdot n^2 \cdot 2^{-(n-1+\ell)/2}\Big)^{n-\ell} \\
&\le \Big(2^{-(\ell-2+\varepsilon_N)+4\log n}\Big)^{(n-\ell)/2} \le 2^{-\Omega(n)}.
\end{aligned} \tag{3.48}$$

To sum up: combining (3.47)–(3.48) implies $\Lambda_1 = \sum_{2 \le \ell < n} w_\ell = o(\mu)$, which completes the proof of (3.3). □

The above calculations yield $\mathrm{Var} X_H = (1+o(1))\mathbb{E} X_H$ by (3.33), and can thus also be used to give an alternative (and compared to Section 3.2.1 more direct) proof of the sharp threshold result Theorem 3.1.1 (i) for $p_2 = 1/2$. We leave the optimal rate of convergence in (3.3) as an intriguing open problem.

## 3.3 Maximum common induced subgraph problem

In this section we prove Theorem 3.1.3, i.e., establish two-point concentration of the size $I_N$ of the maximum common induced subgraph of two independent random graphs $G_{N,p_1}$ and $G_{N,p_2}$ with constant edge-probabilities $p_1, p_2 \in (0,1)$. It naturally splits into two parts: we prove the whp upper bound $I_N \le \lfloor n_N + \varepsilon_N \rfloor$ in Section 3.3.1, and prove the whp lower bound $I_N \ge \lfloor n_N - \varepsilon_N \rfloor$ in Section 3.3.2.

To analyze the parameter $n_N$ defined in (3.11), it will be convenient to study the auxiliary function

$$g(p) := \max\big\{\log b_0(p),\, 2\log b_1(p),\, 2\log b_2(p)\big\} \qquad \text{for } p \in [0,1], \tag{3.49}$$

where the functions $b_0(p)$ and $b_i(p)$ are defined as in (3.7). Note that $g(p)$ depends on $p_1, p_2 \in (0,1)$, but not on $N$. The following key analytic lemma establishes lower bounds on $b_0, b_1$ and $b_2$, as well as properties of $g$ (the proof is based on standard calculus techniques, and thus deferred to Section 3.4).

**Lemma 3.3.1.** *The functions $g(p)$ and $b_j(p)$ with $j \in \{0,1,2\}$ have the following properties:*

*(i) The function $g : [0,1] \to (0,\infty)$ has a unique minimizer $p_0 \in (0,1)$. Furthermore, there is $\xi = \xi(p_1,p_2) \in (0,1/2)$ such that $g(p) \ge g(p_0) + \xi$ for all $p \in [0,\xi] \cup [1-\xi,1]$.*

*(ii) We have $\min_{p\in[0,1]} b_i(p) \ge 1$ for $i \in \{1,2\}$. Furthermore, there is $\lambda = \lambda(p_1,p_2) > 1$ such that $\min_{p\in[0,1]} b_0(p) \ge \lambda$, and for sufficiently large $N \ge N_0(p_1,p_2)$ the following holds for any $p = p(N) \in [0,1]$ and $i \in \{1,2\}$: if $x_N^{(i)}(p) \le x_N^{(0)}(p)$, then $b_i(p) \ge \lambda$. In particular,*

*for any $p = p(N) \in [0,1]$ we have*

$$\min\left\{x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p)\right\} = \frac{4\log N}{g(p)} + O(\log\log N) \tag{3.50}$$

*where the implicit constants in* (3.50) *depend only on $p_1, p_2$ (and not on $p$).*

Note that, using the uniformity of estimate (3.50), now (3.6) and $n_N = \Theta(\log N)$ follow readily from

$$n_N = \max_{p\in[0,1]} \min\left\{x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p)\right\} = \frac{4\log N}{g(p_0)} + O(\log\log N). \tag{3.51}$$

### 3.3.1 Upper bound: No common induced subgraph of size $\lceil n_N + \varepsilon_N \rceil$

In this section we prove the upper bound in (3.5) of Theorem 3.1.3. More precisely, we show that whp there is no common induced subgraph of size $\lceil n_N + \varepsilon_N \rceil$ in $G_{N,p_1}$ and $G_{N,p_2}$, which implies the desired whp upper bound $I_N \le \lceil n_N + \varepsilon_N \rceil - 1 \le \lfloor n_N + \varepsilon_N \rfloor$. Our proof-strategy employs a refinement of the standard first moment method: the idea is to apply different first moment bounds for different densities of the potential common induced subgraphs, which in turn deal with the three containment bottlenecks discussed in Section 3.1.3 (i.e., containment in $G_{N,p_1}$, containment in $G_{N,p_2}$, and containment in both). As we shall see, these estimates are enabled by the corresponding three terms appearing in the definition (3.11) of $n_N$.

Turning to the details, to avoid clutter we define

$$n := \lceil n_N + \varepsilon_N \rceil. \tag{3.52}$$

Writing $\mathscr{B}_m$ for the 'bad' event that $G_{N,p_1}$ and $G_{N,p_2}$ have a common induced subgraph with $n$ vertices and $m$ edges, using a standard union bound argument it follows that

$$\mathbb{P}(I_N \ge \lceil n_N + \varepsilon_N \rceil) \le \sum_{0\le m\le \binom{n}{2}} \mathbb{P}(\mathscr{B}_m). \tag{3.53}$$

Since we are only interested in equality of common subgraphs up to isomorphisms, we define $\mathscr{G}_{n,m}$ as the set of all unlabeled graphs with $n$ vertices and $m$ edges. Let $X_H^{(1)}$ and $X_H^{(2)}$ denote the number of induced copies of $H$ in $G_{N,p_1}$ and $G_{N,p_2}$, respectively. The crux is that if $\mathscr{B}_m$ holds, then the three inequalities $\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)} \geq 1$ and $\sum_{H\in\mathscr{G}_{n,m}} X_H^{(2)} \geq 1$ as well as $\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)} X_H^{(2)} \geq 1$ all hold. Invoking Markov's inequality three times, it thus follows that the bad event $\mathscr{B}_m$ holds with probability at most

$$\mathbb{P}(\mathscr{B}_m) \leq \min\Bigg\{\mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)},\ \mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(2)},\ \mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)} X_H^{(2)}\Bigg\}, \tag{3.54}$$

where the three expectations correspond to the three containment bottlenecks discussed in Section 3.1.3. Let

$$p = p(m,n) := \frac{m}{\binom{n}{2}}. \tag{3.55}$$

The form of (3.54) suggests that we might need a good estimate of $|\mathscr{G}_{n,m}|$, but it turns out that we can avoid this: by double-counting labeled graphs on $n$ vertices with $m$ edges, the crux is that we obtain the identity

$$\sum_{H\in\mathscr{G}_{n,m}} \frac{n!}{|\mathrm{Aut}(H)|} = \binom{\binom{n}{2}}{m}, \tag{3.56}$$

which in view of (3.31) for $G_{N,p_2}$ interacts favorably with the form of the expectations $\mathbb{E}X_H^{(1)}$ and $X_H^{(2)}$. We shall further approximate (3.56) using the following consequence of Stirling's approximation formula:

$$\binom{\binom{n}{2}}{m} = e^{O(\log n)} \left(\frac{1}{p^p(1-p)^{1-p}}\right)^{\binom{n}{2}} \qquad \text{when } m = p\binom{n}{2}. \tag{3.57}$$

The heuristic idea for bounding $\mathbb{P}(\mathscr{B}_m)$ is to focus on the smallest expectation in (3.54) for $m = p\binom{n}{2}$, but due to the definition (3.11) of the parameter $n_N$ in $n = \lceil n_N + \varepsilon_N \rceil$ it will be easier to use a case distinction depending on which term attains the minimum among $x_N^{(0)}(p)$, $x_N^{(1)}(p)$ and $x_N^{(2)}(p)$.

**Case 1.** $x_N^{(i)}(p) = \min\{x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p)\}$ *for* $i \in \{1,2\}$.

Here we focus on $\mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(i)}$ in (3.54). Invoking the estimates (3.56)–(3.57) to bound $\sum_{H\in\mathscr{G}_{n,m}} 1/|\mathrm{Aut}(H)|$, by using the definition (3.7) of $b_i = b_i(p) = (p/p_i)^p[(1-p)/(1-$

$p_i)]^{1-p}$ and $n^2/N = O((\log N)^2/N) = o(1)$ it follows similarly to (3.31) that

$$\begin{aligned}\mathbb{P}(\mathscr{B}_m) &\le \mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(i)} = \sum_{H\in\mathscr{G}_{n,m}} \mathbb{E}X_H^{(i)} \\ &= \sum_{H\in\mathscr{G}_{n,m}} \frac{1}{|\mathrm{Aut}(H)|}\cdot (N)_n \cdot p_i^{p\binom{n}{2}}(1-p_i)^{(1-p)\binom{n}{2}} \\ &= e^{O(\log n)}\frac{N^n}{n!} b_i^{-\binom{n}{2}}.\end{aligned} \tag{3.58}$$

Since $x_N^{(i)} = x_N^{(i)}(p) \le x_N^{(0)}(p)$, by Lemma 3.3.1 (ii) we have $b_i = b_i(p) \ge \lambda$ for some constant $\lambda = \lambda(p_1,p_2) > 1$. Inserting Stirling's approximation formula $n! = e^{O(\log n)}(n/e)^n$ and the identity $eN = x_N^{(i)} b_i^{(x_N^{(i)}-1)/2}$ from (3.12) into estimate (3.58), using $n \ge n_N + \varepsilon_N = x_N^{(i)} + \varepsilon_N$ (which implies $x_N^{(i)} \le n$ and $n - x_N^{(i)} \ge \varepsilon_N$) it follows that

$$\mathbb{P}(\mathscr{B}_m) \le \left[e^{O(\log n/n)}\tfrac{eN}{n} b_i^{-(n-1)/2}\right]^n = \left[e^{O(\log n/n)}\tfrac{x_N^{(i)}}{n} b_i^{-(n-x_N^{(i)})/2}\right]^n \le \lambda^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)}. \tag{3.59}$$

**Case 2.** $x_N^{(0)}(p) = \min\{x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p)\}$.

Here we focus on $\mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)}X_H^{(2)}$ in (3.54). Exploiting independence of the two random graphs $G_{N,p_1}$ and $G_{N,p_2}$, using $|\mathrm{Aut}(H)|^2 \ge |\mathrm{Aut}(H)|$, and applying the definition (3.7) of $b_0 = b_0(p)$ it follows similarly to (3.58) that

$$\begin{aligned}\mathbb{P}(\mathscr{B}_m) &\le \mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)}X_H^{(2)} = \sum_{H\in\mathscr{G}_{n,m}} \mathbb{E}X_H^{(1)}\mathbb{E}X_H^{(2)} \\ &= \sum_{H\in\mathscr{G}_{n,m}} \frac{1}{|\mathrm{Aut}(H)|^2}\cdot (N)_n^2\cdot \big(p_1p_2\big)^{p\binom{n}{2}}\big((1-p_1)(1-p_2)\big)^{(1-p)\binom{n}{2}} \\ &\le e^{O(\log n)}\frac{N^{2n}}{n!} b_0{}^{-\binom{n}{2}}.\end{aligned} \tag{3.60}$$

By Lemma 3.3.1 (ii) we have $b_0 = b_0(p) \ge \lambda$ for some constant $\lambda = \lambda(p_1,p_2) > 1$. Recalling that $\log N = \Theta(n)$ by (3.51), observe that the definition (3.10) of $x_N^{(0)} = x_N^{(0)}(p)$ ensures that

$$x_N^{(0)} b_0^{(x_N^{(0)}-1)/2} = \tfrac{x_N^{(0)}}{4\log_{b_0}(N)}\cdot eN^2 = \Big(1 + O\big(\tfrac{\log\log N}{\log N}\big)\Big)\cdot eN^2 = e^{O((\log n)/n)} eN^2. \tag{3.61}$$

Inserting Stirling's approximation formula $n! = e^{O(\log n)}(n/e)^n$ and (3.61) into (3.60), using $n \ge$

$n_N+\varepsilon_N = x_N^{(0)}+\varepsilon_N$ (which implies $x_N^{(0)} \le n$ and $n-x_N^{(0)} \ge \varepsilon_N$) it follows that

$$\mathbb{P}(\mathscr{B}_m) \le \left[e^{O(\log n/n)} \tfrac{eN^2}{n} {b_0}^{-(n-1)/2}\right]^n = \left[e^{O(\log n/n)} \tfrac{x_N^{(0)}}{n} b_0^{-(n-x_N^{(0)})/2}\right]^n \le \lambda^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)}. \tag{3.62}$$

To sum up: inserting (3.59) and (3.62) into (3.53) readily gives $\mathbb{P}(I_N \ge \lceil n_N+\varepsilon_N\rceil) = o(1)$, which as discussed completes the proof of the upper bound in (3.5), i.e., that whp $I_N \le \lfloor n_N+\varepsilon_N\rfloor$.

### 3.3.2 Lower bound: Common induced subgraph of size $\lfloor n_N-\varepsilon_N\rfloor$

In this section we prove the lower bound in (3.5) of Theorem 3.1.3. More precisely, we establish the desired whp lower bound $I_N \ge \lfloor n_N-\varepsilon_N\rfloor$ by showing that whp there is a common induced subgraph of size $\lfloor n_N-\varepsilon_N\rfloor$ in $G_{N,p_1}$ and $G_{N,p_2}$. Our proof-strategy is inspired by that of Lemma 3.2.1 from Section 3.2.1, though the technical details are significantly more involved: the idea is to pick an 'optimal' edge-density $p$, and then apply the second moment method to the total number of pairs of induced copies of $H$ in $G_{N,p_1}$ and $G_{N,p_2}$, where we consider only pseudorandom graphs $H$ with $\lfloor n_N-\varepsilon_N\rfloor$ vertices and $\lfloor p\binom{n}{2}\rfloor$ edges. Here the restriction to pseudorandom $H$ will again be key for controlling the expectation and variance, with the extra wrinkle that the resulting involved variance calculations share some similarities with fourth moment arguments (that require some new ideas to control the 'overlaps' of different $H$, including more careful enumeration arguments).

Turning to the details, we pick $p$ as a maximizer of $\min\{x_N^{(0)}(p), x_N^{(1)}(p), x_N^{(2)}(p)\}$. By comparing the asymptotic estimate (3.50) with the asymptotics (3.51) of $n_N$, it follows from Lemma 3.3.1 that there is a constant $\xi=\xi(p_1,p_2)\in(0,1/2)$ such that $p\in[\xi,1-\xi]$ for all sufficiently large $N$ (otherwise the first order asymptotics of (3.50) and (3.51) would differ, contradicting our choice of $p$). To avoid clutter, we define

$$n := \lfloor n_N-\varepsilon_N\rfloor \qquad \text{and} \qquad m := \left\lfloor p\tbinom{n}{2}\right\rfloor. \tag{3.63}$$

Recall that in Section 3.2.1 we introduced the set $\mathscr{T}_{n,m} = \mathscr{A}_n \cap \mathscr{E}_{n,m}$ of pseudorandom graphs with $n$ vertices and $m$ edges, where $\mathscr{A}_n$ and $\mathscr{E}_{n,m}$ are defined as in Section 3.1.3. Since we are only interested in the existence of isomorphisms between induced subgraphs, we now define $\mathscr{T}$ as the unlabeled variant of $\mathscr{T}_{n,m}$, which can formally be constructed by ignoring labels of the graphs in $\mathscr{T}_{n,m}$. Since any graph in $\mathscr{T}_{n,m} \subseteq \mathscr{A}_n$ is asymmetric, we have $|\mathscr{T}| = |\mathscr{T}_{n,m}|/n!$, and so it follows from Lemma 3.1.6 that

$$|\mathscr{T}| = (1+o(1))\binom{\binom{n}{2}}{m}\frac{1}{n!}. \tag{3.64}$$

As in Section 3.3.1, let $X_H^{(1)}$ and $X_H^{(2)}$ denote the number of induced copies of $H$ in $G_{N,p_1}$ and $G_{N,p_2}$, respectively. We then define the random variable

$$X := \sum_{H\in\mathscr{T}} X_H^{(1)}X_H^{(2)}, \tag{3.65}$$

where $X > 0$ implies that $G_{N,p_1}$ and $G_{N,p_2}$ have a common induced subgraph on $n$ vertices, i.e., that $I_N \geq n$. To complete the proof of $\mathbb{P}(I_N \geq \lfloor n_N - \varepsilon_N \rfloor) = 1 - o(1)$, using Chebyshev's inequality it thus suffices to show that $\mathbb{E}X \to \infty$ and $\mathrm{Var}X = o\big((\mathbb{E}X)^2\big)$ as $N \to \infty$.

We start by showing that $\mathbb{E}X \to \infty$ as $N \to \infty$. Analogous to Section 3.2.1 we have $X_H^{(i)} = \sum_{R_i\in\binom{[N]}{n}} I_{H,R_i}^{(i)}$, where $I_{H,R_i}^{(i)}$ is the indicator random variable for the event that the induced subgraph $G_{N,p_i}[R_i]$ is isomorphic to $H$. Note that every unlabeled graph $H \in \mathscr{T}$ satisfies the asymmetry property $\mathscr{A}_{n,m}$ from Section 3.1.3, so that $|\mathrm{Aut}(H)| = 1$. It thus follows similarly to (3.60) in Section 3.3.1 that

$$\mu := \mathbb{E}X = \sum_{H\in\mathscr{T}} \mathbb{E}X_H^{(1)}\mathbb{E}X_H^{(2)} = |\mathscr{T}| \cdot \tbinom{N}{n}^2 \cdot \mu_1\mu_2, \tag{3.66}$$

where

$$\mu_i := \mathbb{E}I_{H,R_i}^{(i)} = n!p_i^m(1-p_i)^{\binom{n}{2}-m}. \tag{3.67}$$

Inspecting the form of (3.66) and (3.56), note that the asymptotic estimate (3.64) of $|\mathscr{T}|$ allows us to estimate $\mathbb{E}X$ analogously to $\mathbb{E}\sum_{H\in\mathscr{G}_{n,m}} X_H^{(1)}X_H^{(2)}$ in (3.60)–(3.62). Indeed, using $b_0 = b_0(p) \geq \lambda = \lambda(p_1,p_2) > 1$ and $n \leq n_N - \varepsilon_N \leq x_N^{(0)} - \varepsilon_N$ (which implies $x_N^{(0)} \geq n$ and

$x_N^{(0)} - n \geq \varepsilon_N$) it here follows that

$$\begin{aligned}\mathbb{E}X &= (1+o(1))\binom{\binom{n}{2}}{m}\frac{1}{n!}\cdot (N)_n^2 \big(p_1p_2\big)^{p\binom{n}{2}}\big((1-p_1)(1-p_2)\big)^{(1-p)\binom{n}{2}} \\ &= e^{O(\log n)}\frac{N^{2n}}{n!}b_0{}^{-\binom{n}{2}} = \left[e^{O(\log n/n)}\frac{x_N^{(0)}}{n}b_0^{(x_N^{(0)}-n)/2}\right]^n \geq \lambda^{\Omega(\varepsilon_N n)} \geq n^{\Omega(\log n)} \to \infty.\end{aligned} \tag{3.68}$$

The remainder of this section is devoted to showing that $\operatorname{Var} X = o\big((\mathbb{E}X)^2\big)$. Note that $I_{H,R_1}^{(1)}$ and $I_{H',R_2}^{(2)}$ are always independent. Since $I_{H,R_i}^{(i)}$ and $I_{H',S_i}^{(i)}$ are independent when $|R_i \cap S_i| \leq 1$, it follows that

$$\operatorname{Var} X \leq \sum_{\substack{0\leq \ell_1,\ell_2\leq n:\\ \max\{\ell_1,\ell_2\}\geq 2}} \underbrace{\sum_{H,H'\in\mathscr{T}}\Bigg(\sum_{\substack{R_1,S_1\in\binom{[N]}{n}:\\ |R_1\cap S_1|=\ell_1}} \mathbb{E}I_{H,R_1}^{(1)}I_{H',S_1}^{(1)}\Bigg)\Bigg(\sum_{\substack{R_2,S_2\in\binom{[N]}{n}:\\ |R_2\cap S_2|=\ell_2}} \mathbb{E}I_{H,R_2}^{(2)}I_{H',S_2}^{(2)}\Bigg)}_{=:w_{\ell_1,\ell_2}}. \tag{3.69}$$

Our upcoming estimates of this somewhat elaborate variance expression use a case distinction depending on whether $\ell_1 \geq \max\{\ell_2, 2\}$ or $\ell_2 \geq \max\{\ell_1, 2\}$ holds. Both cases can be handled by the same argument (with the roles of $G_{N,p_1}$ and $G_{N,p_2}$ in the below definition (3.70) of $u_{\ell_1}$ and $v_{\ell_2}$ interchanged), so we shall henceforth focus on the case where $\ell_1 \geq \max\{\ell_2, 2\}$ holds. In this case we find it convenient to estimate

$$\frac{w_{\ell_1,\ell_2}}{\mu^2} \leq \underbrace{\frac{\sum_{H,H'\in\mathscr{T}}\sum_{R_1,S_1:|R_1\cap S_1|=\ell_1}\mathbb{E}I_{H,R_1}^{(1)}I_{H',S_1}^{(1)}}{|\mathscr{T}|^2\binom{N}{n}^2\mu_1^2}}_{=:u_{\ell_1}} \cdot \underbrace{\max_{H,H'\in\mathscr{T}}\frac{\sum_{R_2,S_2:|R_2\cap S_2|=\ell_2}\mathbb{E}I_{H,R_2}^{(2)}I_{H',S_2}^{(2)}}{\binom{N}{n}^2\mu_2^2}}_{=:v_{\ell_2}}, \tag{3.70}$$

where the sums are taken over all vertex sets $R_1, S_1 \in \binom{[N]}{n}$ and $R_2, S_2 \in \binom{[N]}{n}$, as in (3.69) above. This splitting allows us to deal with one random graph $G_{N,p_i}$ at a time, which we shall exploit when bounding $u_{\ell_1}$ and $v_{\ell_2}$ in the upcoming Sections 3.3.2–3.3.2. For later reference, we now define the (sufficiently small) constant

$$\zeta := \min\left\{\frac{\log b_0}{4\log\max\left\{\frac{1}{p_1},\frac{1}{1-p_1},\frac{1}{p_2},\frac{1}{1-p_2}\right\}}, \frac{1}{2}\right\}. \tag{3.71}$$

**Contribution of $G_{N,p_2}$ to variance: Bounding $v_\ell$**

We first bound the parameter $v_\ell$ defined in (3.70) for $0 \le \ell \le n$, using similar arguments as for the variance calculations from Section 3.2.1. We start with the pathological case $0 \le \ell \le 1$, where in view of (3.67) we trivially (due to independence of $I^{(2)}_{H,R_2}$ and $I^{(2)}_{H',S_2}$ when $|R_2 \cap S_2| \le 1$) have

$$v_\ell = \max_{H,H' \in \mathscr{T}} \frac{\sum_{R_2,S_2:|R_2\cap S_2|=\ell} \mathbb{E}I^{(2)}_{H,R_2}\mathbb{E}I^{(2)}_{H',S_2}}{\binom{N}{n}^2 \mu_2^2} \le 1 \qquad \text{when } 0 \le \ell \le 1. \tag{3.72}$$

It remains to bound $v_\ell$ for $2 \le \ell \le n$. Here we shall reuse some ideas from Section 3.2.1: bounding the numerator in the definition (3.70) of $v_\ell$ as in (3.34)–(3.35), in view of $\mu_2 = n! p_2^m (1-p_2)^{\binom{n}{2}-m}$ it follows that

$$\begin{aligned} v_\ell &\le \frac{\binom{N}{n}\binom{n}{\ell}\binom{N-n}{n-\ell}\cdot n! \cdot (n)_{n-\ell}}{\binom{N}{n}^2 (n!)^2} \cdot \max_{H\in\mathscr{T},\,|L|=\ell} |\mathrm{Aut}(H[L])| \\ &\qquad\qquad \cdot \max_{H\in\mathscr{T},\,|L|=\ell} \frac{p_2^{2m-e(H[L])}(1-p_2)^{2\binom{n}{2}-2m-\binom{\ell}{2}+e(H[L])}}{p_2^{2m}(1-p_2)^{2\binom{n}{2}-2m}} \\ &\le \underbrace{\frac{(N-n)_{n-\ell}}{(N)_n}}_{=(1+o(1))N^{-\ell}} \cdot \binom{n}{\ell}^2 \cdot \max_{H\in\mathscr{T},\,|L|=\ell} |\mathrm{Aut}(H[L])| \cdot \underbrace{\max_{H\in\mathscr{T},\,|L|=\ell} p_2^{-e(H[L])}(1-p_2)^{-\binom{\ell}{2}+e(H[L])}}_{=:P_{\ell,2}}, \end{aligned} \tag{3.73}$$

where the four maxima in (3.73) are each taken over all vertex-subsets $L \subseteq V(H)$ of size $|L| = \ell$, as before. We define the parameter $P_{\ell,1}$ analogously to $P_{\ell,2}$ (by replacing $p_2$ with $p_1$). Since every unlabeled graph $H \in \mathscr{T}$ satisfies the pseudorandom edge-properties of $\mathscr{E}_{n,m}$ from Section 3.1.3, it follows that

$$P_{\ell,i} \le \max_{k:\left|k-\binom{\ell}{2}m/\binom{n}{2}\right| \le n^{2/3}(n-\ell)} p_i^{-k}(1-p_i)^{-\binom{\ell}{2}+k} \tag{3.74}$$

for any $0 \le \ell \le n$, which in view of $m = \left\lfloor p\binom{n}{2}\right\rfloor$ yields, similarly to (3.36) and (3.38) from

Section 3.2.1, that

$$P_{\ell,i} \le \begin{cases} \left[\max\left\{\frac{1}{p_i}, \frac{1}{1-p_i}\right\}\right]^{\binom{\ell}{2}} & \text{if } 0 \le \ell \le \zeta n, \\ e^{O\left(n^{2/3}(n-\ell)+1\right)} \cdot \left(p_i^p(1-p_i)^{1-p}\right)^{-\binom{\ell}{2}} & \text{if } \zeta n \le \ell \le n. \end{cases} \tag{3.75}$$

After these preparations, we are now ready to bound $v_\ell$ further using a case distinction.

**Case 1.** $2 \le \ell \le \zeta n$.

Here we proceed similarly to (3.36)–(3.37), and exploit that for all large enough $N$ we have

$$\ell \le \zeta n \le \zeta \cdot n_N \le \zeta \cdot x_N^{(0)}(p) \le \zeta \cdot \frac{4\log N}{\log b_0} \le \frac{\log N}{\log\max\left\{\frac{1}{p_1}, \frac{1}{1-p_1}, \frac{1}{p_2}, \frac{1}{1-p_2}\right\}} \tag{3.76}$$

by our choice of $\zeta$. Inserting (3.75) and the trivial bound $|\mathrm{Aut}(H[L])| \le |L|! = \ell!$ into (3.73), using $\binom{n}{\ell} \le n^\ell/\ell!$ and (3.76) it follows for all large enough $N$ that

$$v_\ell \le 2N^{-\ell} \cdot \binom{n}{\ell}^2 \ell! \cdot P_{\ell,2} \le \left(N^{-1}n^2 \cdot \left[\max\left\{\frac{1}{p_2}, \frac{1}{1-p_2}\right\}\right]^{(\ell-1)/2}\right)^\ell \le N^{-\ell/3} \le e^{-\Omega(n)}. \tag{3.77}$$

**Case 2.** $\zeta n \le \ell \le n$.

Here we refine the previous argument, proceeding similarly to (3.39). For $\zeta n \le \ell < n - n^{2/3}$ we estimate $|\mathrm{Aut}(H[L])| \le \ell! \le n^\ell \le e^{O(n^{2/3}(n-\ell))}$, and for $n - n^{2/3} \le \ell \le n$ we have $|\mathrm{Aut}(H[L])| = 1$ since every unlabeled graph $H \in \mathscr{T}$ satisfies the asymmetry property $\mathscr{A}_{n,m}$ from Section 3.1.3. Inserting these estimates and $\binom{n}{\ell} = \binom{n}{n-\ell} \le n^{n-\ell} = e^{O(n^{2/3}(n-\ell))}$ into (3.73), using (3.75) it follows that

$$v_\ell \le 2N^{-\ell} \cdot \binom{n}{\ell}^2 \cdot e^{O\left(n^{2/3}(n-\ell)+1\right)} \cdot P_{\ell,2} \le O(1) \cdot \left[N^{-1}(p_2^p(1-p_2)^{1-p})^{-(\ell-1)/2} e^{O\left(n^{-1/3}(n-\ell)\right)}\right]^\ell, \tag{3.78}$$

which in Section 3.3.2 will later turn out to be a useful upper bound.

**Contribution of $G_{N,p_1}$ to variance: Bounding $u_\ell$**

We now bound the parameter $u_\ell$ defined in (3.70) for $2 \le \ell \le n$. For the simpler case $2 \le \ell \le \zeta n$ we shall reuse the argument from Section 3.3.2, while the more elaborate case $\zeta n \le \ell \le n$ requires further new ideas.

**Case 1.** $2 \le \ell \le \zeta n$.

Using the pair $H, H' \in \mathscr{T}$ which maximizes the summand as an upper bound reduces this case to the analogous bounds for $v_\ell$ from Section 3.3.2. Indeed, by proceeding this way we obtain that

$$u_\ell \le \max_{H,H' \in \mathscr{T}} \frac{\sum_{R_1,S_1:|R_1\cap S_1|=\ell} \mathbb{E} I^{(1)}_{H,R_1} I^{(1)}_{H',S_1}}{\binom{N}{n}^2 \mu_1^2} \le e^{-\Omega(n)}, \tag{3.79}$$

where the last inequality follows word-by-word (with $p_2$ replaced by $p_1$, and $P_{\ell,2}$ replaced by $P_{\ell,1}$) from the arguments leading to (3.76)–(3.77).

**Case 2.** $\zeta n \le \ell \le n$.

Here one key idea is to interchange the order of summation in the numerator to obtain

$$u_\ell = \frac{\sum_{R_1,S_1:|R_1\cap S_1|=\ell} \sum_{H,H' \in \mathscr{T}} \mathbb{E} I^{(1)}_{H,R_1} I^{(1)}_{H',S_1}}{\binom{N}{n}^2 |\mathscr{T}|^2 \mu_1^2}, \tag{3.80}$$

which then allows us to exploit that not too many choices of the pseudorandom graphs $H, H' \in \mathscr{T}$ can intersect in a 'compatible' way. Indeed, if we proceeded similarly to the argument leading to (3.73), (3.77) and (3.79), then after choosing $R_1, S_1 \in \binom{[N]}{n}$ with $|R_1 \cap S_1| = \ell$, in the numerator of (3.80) we would use

$$|\mathscr{T}|^2 \cdot n! \cdot (n)_{n-\ell} \cdot \max_{H \in \mathscr{T},\, |L|=\ell} |\mathrm{Aut}(H[L])| \le |\mathscr{T}|^2 (n!)^2$$

as a simple upper bound on the number $\Lambda_\ell(R_1, S_1)$ of labeled graphs $F$ on $R_1 \cup S_1$ such that $F[R_1]$ and $F[S_1]$ are isomorphic to some $H$ and $H'$, respectively, where $H, H' \in \mathscr{T}$. Here our key improvement idea is to more carefully enumerate all possible such graphs $F$, by first choosing the edges in the intersection $R_1 \cap S_1$ of size $|R_1 \cap S_1| = \ell$, and only then the remaining edges of $F$.

The crux is that since all possible $H, H' \in \mathscr{T}$ satisfy the pseudorandom edge-properties of $\mathscr{E}_{n,m}$ from Section 3.1.3, we know in advance that all possible numbers $k$ of edges inside $R_1 \cap S_1$ must satisfy $\left|k - \binom{\ell}{2}m/\binom{n}{2}\right| \le n^{2/3}(n-\ell)$. Hence, by first choosing $k$ edges in the intersection $R_1 \cap S_1$, and then the remaining $2(m-k)$ edges, using $m = \left\lfloor p\binom{n}{2}\right\rfloor$ it follows that

$$
\begin{aligned}
\Lambda_\ell(R_1, S_1) &\le \sum_{k:\left|k-\binom{\ell}{2}m/\binom{n}{2}\right| \le n^{2/3}(n-\ell)} \binom{\binom{\ell}{2}}{k}\binom{2\left[\binom{n}{2}-\binom{\ell}{2}\right]}{2(m-k)} \\
&\le n^2 \cdot \binom{\binom{\ell}{2}}{\left\lfloor p\binom{\ell}{2}\right\rfloor}\binom{2\left[\binom{n}{2}-\binom{\ell}{2}\right]}{\left\lfloor 2p\left[\binom{n}{2}-\binom{\ell}{2}\right]\right\rfloor} \cdot n^{O\left(n^{2/3}(n-\ell)+1\right)} \qquad (3.81) \\
&\le \left(p^p(1-p)^{1-p}\right)^{-2\binom{n}{2}+\binom{\ell}{2}} \cdot e^{O\left(\log n + n^{3/4}(n-\ell)\right)},
\end{aligned}
$$

where for the last inequality we used Stirling's approximation formula similarly to (3.57). Using again Stirling's approximation formula similarly to (3.57), from the asymptotic estimate (3.64) for $|\mathscr{T}|$ it follows that

$$
|\mathscr{T}|^2(n!)^2 = (1+o(1))\binom{\binom{n}{2}}{m}^2 = e^{O(\log n)}\left(p^p(1-p)^{1-p}\right)^{-2\binom{n}{2}}, \qquad (3.82)
$$

which in view of $\binom{\ell}{2} = \Theta(n^2)$ makes it transparent that the refined upper bound (3.81) on $\Lambda_\ell(R_1, S_1)$ is significantly smaller than the simple upper bound $|\mathscr{T}|^2(n!)^2$ mentioned above.

After these preparations, we are now ready to estimate $u_\ell$ as written in (3.80): namely, (i) we bound the numerator as in (3.73), the key difference being that we use $\Lambda_\ell(R_1, S_1)$ to account for the choices of $H, H' \in \mathscr{T}$ and their embeddings into $R_1$ and $S_1$, and (ii) we also use (3.82) to estimate the denominator in (3.80). Taking these differences to (3.73) into account, using $\mu_1 = n!p_1^m(1-p_1)^{\binom{n}{2}-m}$ it then follows that

$$
\begin{aligned}
u_\ell &\le \frac{\binom{N}{n}\binom{n}{\ell}\binom{N-n}{n-\ell}}{\binom{N}{n}^2|\mathscr{T}|^2(n!)^2} \cdot \max_{R_1,S_1:|R_1\cap S_1|=\ell} \Lambda_\ell(R_1,S_1) \\
&\qquad \cdot \max_{H\in\mathscr{T},\,|L|=\ell} \frac{p_1^{2m-e(H[L])}(1-p_1)^{2\binom{n}{2}-2m-\binom{\ell}{2}+e(H[L])}}{p_1^{2m}(1-p_1)^{2\binom{n}{2}-2m}} \\
&\le (1+o(1))N^{-\ell} \cdot \binom{n}{\ell}^2 \ell! \cdot \left(p^p(1-p)^{1-p}\right)^{\binom{\ell}{2}} e^{O\left(\log n + n^{3/4}(n-\ell)\right)} \cdot P_{\ell,1},
\end{aligned}
$$

where $P_{\ell,1}$ is defined as below (3.73). Combining the upper bound (3.74)–(3.75) for $P_{\ell,1}$ with the definition (3.7) of $b_1 = b_1(p) = (p/p_1)^p[(1-p)/(1-p_1)]^{1-p}$, using Stirling's approximation formula $\ell! = e^{O(\log n)}(\ell/e)^\ell$ and $\binom{n}{\ell}^2 = \binom{n}{n-\ell}^2 \le n^{2(n-\ell)} = e^{O(n^{3/4}(n-\ell))}$ it follows that

$$u_\ell \le \left(\tfrac{\ell}{eN}\right)^\ell \cdot b_1^{\binom{\ell}{2}} \cdot e^{O\left(\log n + n^{3/4}(n-\ell)\right)} \le \left[e^{O\left(\log n/n + n^{-1/4}(n-\ell)\right)} \tfrac{\ell}{eN} b_1^{(\ell-1)/2}\right]^\ell. \tag{3.83}$$

We now find it convenient to treat the case where $b_1$ is close to 1 separately from the case where $b_1$ is bounded away from 1. From (3.51) we know that $\ell \le n = O(\log N)$, so there exists a constant $\rho = \rho(p_1, p_2) > 0$ such that $b_1 \le 1+\rho$ implies $b_1^{(\ell-1)/2} \le e^{\rho(\ell-1)/2} \le \sqrt{N}$. In case of $b_1 \le 1+\rho$ we thus infer that

$$u_\ell \le \left[e^{O(n^{3/4})} \tfrac{\ell}{eN} b_1^{(\ell-1)/2}\right]^\ell \le \left[e^{O(n^{3/4})} \tfrac{\ell}{\sqrt{N}}\right]^\ell \le n^{-\Omega(\log n)}. \tag{3.84}$$

In case of $b_1 \ge 1+\rho$ we insert the identity $eN = x_N^{(1)} b_1^{(x_N^{(1)}-1)/2}$ from (3.12) into estimate (3.83), and then use $\zeta n \le \ell \le n \le x_N^{(1)}$ (which implies $0 \le n-\ell \le x_N^{(1)} - \ell$ and $\ell \le x_N^{(1)}$ as well as $x_N^{(1)} - \ell \ge x_N^{(1)} - n \ge \varepsilon_N$ and $\ell \ge \zeta n$) to infer that

$$\begin{aligned} u_\ell &\le \left[e^{O(\log n/n) + o(n-\ell)} \tfrac{\ell}{x_N^{(1)}} b_1^{(\ell - x_N^{(1)})/2}\right]^\ell \\ &\le \left[e^{O(\log n/n)} (1+\rho)^{-(x_N^{(1)}-\ell)/3}\right]^\ell \\ &\le e^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)}. \end{aligned} \tag{3.85}$$

To sum up: from the two estimates (3.84)–(3.85) it follows that we always have

$$u_\ell \le n^{-\Omega(\log n)} \qquad \text{when } \zeta n \le \ell \le n. \tag{3.86}$$

**Putting things together: Bounding $w_{\ell_1,\ell_2}/\mu^2$**

Using the estimates from Sections 3.3.2–3.3.2, in the case $\ell_1 \ge \max\{\ell_2, 2\}$ we are now ready to bound the parameter $w_{\ell_1,\ell_2}/\mu^2 \le u_{\ell_1} v_{\ell_2}$ from (3.70) using a case distinction based on whether or not $\ell_2 \le \zeta n$.

**Case 1.** $0 \le \ell_2 \le \zeta n$ *and* $2 \le \ell_1 \le n$.

From (3.72) and (3.77), we infer that $v_{\ell_2} \le 1$ for all $0 \le \ell_2 \le \zeta n$. From (3.79) and (3.86), we infer that $u_{\ell_1} \le n^{-\Omega(\log n)}$ for all $2 \le \ell_1 \le n$. Hence

$$u_{\ell_1} v_{\ell_2} \le n^{-\Omega(\log n)}. \tag{3.87}$$

**Case 2.** $\zeta n \le \ell_2 \le \ell_1 \le n$.

Here we split our analysis into two cases, based on whether or not the upper bound (3.78) on $v_{\ell_2}$ is effective on its own. We start with the case where $N(p_2^p(1-p_2)^{1-p})^{(\ell_2-1)/2} \ge e^{n^{3/4}}$ holds: here estimate (3.78) implies $v_{\ell_2} \le e^{-\Omega(\ell_2 n^{3/4})} = o(1)$, which together with (3.86) yields that

$$u_{\ell_1} v_{\ell_2} \le n^{-\Omega(\log n)}. \tag{3.88}$$

We henceforth consider the remaining case, where in view of $\ell_1 = \Omega(n)$ we have

$$N(p_2^p(1-p_2)^{1-p})^{(\ell_1+\ell_2-1)/2} \le e^{n^{3/4}} \cdot (p_2^p(1-p_2)^{1-p})^{\ell_1/2} \le e^{-\Omega(n)}.$$

By combining the estimates (3.83) and (3.78) for $u_{\ell_1}$ and $v_{\ell_2}$ with

$$\binom{\ell_2}{2} = \binom{\ell_1}{2} - (\ell_1-\ell_2)(\ell_1+\ell_2-1)/2$$

as well as $b_0 = b_1 \cdot p_2^{-p}(1-p_2)^{-(1-p)}$ and the identity (3.61), in view of $\ell_2 = O(n)$ and $\ell_1 = \Omega(n)$ it follows that

$$\begin{aligned} u_{\ell_1} v_{\ell_2} &\le \left(\left(\tfrac{\ell_1}{eN}\right)^{\ell_1} b_1^{\binom{\ell_1}{2}}\right) \cdot \left(N^{-\ell_2}(p_2^p(1-p_2)^{1-p})^{-\binom{\ell_2}{2}}\right) \cdot e^{O\left(\log n + n^{3/4}(n-\ell_2)\right)} \\ &\le \left[\tfrac{\ell_1}{eN^2} b_0^{(\ell_1-1)/2}\right]^{\ell_1} \cdot \left[N(p_2^p(1-p_2)^{1-p})^{(\ell_1+\ell_2-1)/2}\right]^{\ell_1-\ell_2} \cdot e^{O\left(\log n + n^{3/4}(n-\ell_1) + n^{3/4}(\ell_1-\ell_2)\right)} \\ &\le \left[e^{O(\log n/n)+o(n-\ell_1)} \frac{\ell_1}{x_N^{(0)}} b_0^{-(x_N^{(0)}-\ell_1)/2}\right]^{\ell_1} \cdot e^{-\Omega(n(\ell_1-\ell_2))}. \end{aligned}$$

Recall that by Lemma 3.3.1 (ii) we have $b_0 = b_0(p) \ge \lambda$ for some constant $\lambda = \lambda(p_1, p_2) > 1$. Using $\zeta n \le \ell_1 \le n \le n_N - \varepsilon_N \le x_N^{(0)} - \varepsilon_N$ (which implies $0 \le n - \ell_1 \le x_N^{(0)} - \ell_1$ and $\ell_1 \le x_N^{(0)}$ as well as $x_N^{(0)} - \ell_1 \ge \varepsilon_N$ and $\ell_1 \ge \zeta n$) it follows that

$$u_{\ell_1} v_{\ell_2} \le \left[e^{O(\log n/n)} \lambda^{-(x_N^{(0)}-\ell_1)/3}\right]^{\ell_1} \le e^{-\Omega(\varepsilon_N n)} \le n^{-\Omega(\log n)}. \tag{3.89}$$

To sum up: in the case $\ell_1 \ge \max\{\ell_2, 2\}$ we each time obtained $w_{\ell_1,\ell_2}/\mu^2 \le n^{-\Omega(\log n)}$.

The same argument (with the roles of $G_{N,p_1}$ and $G_{N,p_2}$ interchanged in the definition (3.70) of $u_{\ell_1}$ and $v_{\ell_2}$) also gives $w_{\ell_1,\ell_2}/\mu^2 \le n^{-\Omega(\log n)}$ when $\ell_2 \ge \max\{\ell_1, 2\}$. Inserting these estimates into (3.69) readily implies $\operatorname{Var} X = o((\mathbb{E}X)^2)$. As discussed, this completes the proof of the lower bound in (3.5), i.e., that whp $I_N \ge \lfloor n_N - \varepsilon_N \rfloor$, which together with the upper bound from Section 3.3.1 completes the proof of Theorem 3.1.3. □

From Theorem 3.1.3 it readily follows that there is a set $M \subseteq \mathbb{N}$ with density 1 such that, for any $N \in M$, we have $\lfloor n_N - \varepsilon_N \rfloor = \lfloor n_N + \varepsilon_N \rfloor$ and thus whp $I_N = \lfloor n_N - \varepsilon_N \rfloor$. We leave it as an interesting open problem to show, that for infinitely many $N$, the size $I_N$ can be equal to the each of the two different numbers $\lfloor n_N - \varepsilon_N \rfloor$ and $\lfloor n_N + \varepsilon_N \rfloor$ with probabilities bounded away from 0 (similar as for cliques in $G_{N,p}$; see [Bol01, Theorem 11.7]).

## 3.4 Locating the parameter $n_N$ from Theorem 3.1.3

In this section we approximately determine the value of the parameter $n_N$ from Theorem 3.1.3, which in (3.11) of Section 3.1.2 is defined as the solution to an optimization problem over all edge-densities $p \in [0,1]$. In particular, Lemma 3.4.1 below locates $n_N$ in terms of the unique minimizer $p_0$ of the auxiliary function

$$g(p) = g_{p_1,p_2}(p) := \max\big\{\log b_0(p),\, 2\log b_1(p),\, 2\log b_2(p)\big\} \qquad \text{for } p \in [0,1], \tag{3.90}$$

where the functions $b_0(p)$ and $b_i(p)$ that depend on $p_1, p_2$ are defined as in (3.7). Using the standard convention that $0\log 0 = \lim_{x\searrow 0} x\log x = 0$ (which is consistent with $0^0 = 1$), for $i \in \{1,2\}$ we here continuously extend $\log b_i(p)$ to $p \in \{0,1\}$, as usual. as in Section 3.1.2. We also introduce the parameter

$$\hat{p} = \hat{p}(p_1, p_2) := \frac{p_1 p_2}{p_1 p_2 + (1-p_1)(1-p_2)} \tag{3.91}$$

that occurs in the different cases of Lemma 3.4.1, which each assume information about the form of $g(\hat{p})$. In (3.92)–(3.93) below we use $a_N \sim b_N$ as a shorthand for $a_N = (1+o(1))b_N$ as $N \to \infty$, as usual.

**Lemma 3.4.1** (Locating the parameter $n_N$). *Fix $p_1, p_2 \in (0,1)$. Then the function $g = g_{p_1,p_2} : [0,1] \to (0,\infty)$ from* (3.90) *is strictly convex, and has a unique minimizer $p_0 \in (0,1)$. Furthermore, writing $\hat{p} = \hat{p}(p_1, p_2)$ as in* (3.91)*, the following holds for $p_0$ and the parameter $n_N$ from Theorem 3.1.3 defined in* (3.11)*:*

*(i) If $\log b_0(\hat{p}) > \max\{2\log b_1(\hat{p}), 2\log b_2(\hat{p})\}$, then $p_0 = \hat{p}$. Moreover, for all large enough $N$, we have*

$$n_N = x_N^{(0)}(\hat{p}) \sim 4\log_{b_0(\hat{p})} N. \tag{3.92}$$

*(ii) If $\log b_0(\hat{p}) \leq 2\log b_i(\hat{p})$ for $i \in \{1,2\}$, then $p_0$ is the unique solution of $\log b_0(p) = 2\log b_i(p)$, and $p_0$ lies between $\hat{p}$ and $p_i$, with $p_0 \neq p_i$. Moreover, for all large enough $N$, we have*

$$n_N = x_N^{(i)}(p_0) + O(\log\log N) \sim 2\log_{b_i(p_0)} N. \tag{3.93}$$

The proof of Lemma 3.4.1 is based on standard calculus techniques (mainly using convexity), and is spread across the following subsections. As a byproduct of these techniques, in Section 3.4.1 and 3.4.2 we also give the deferred proofs of the closely related results Corollary 3.1.4, Remark 3.1.5 and Lemma 3.3.1.

### 3.4.1 Proofs of Remark 3.1.5 and Lemma 3.3.1

The functions $g(p)$ and $\log b_j(p)$ appearing in the following auxiliary lemma of course depend on $p_1, p_2 \in (0,1)$, but to avoid clutter we henceforth suppress the dependence on $p_1, p_2$ in the notation, as usual.

**Lemma 3.4.2.** *Fix $p_1, p_2 \in (0,1)$. The functions $\log b_0(p)$, $\log b_1(p)$ and $\log b_2(p)$ are strictly convex functions for $p \in [0,1]$, and achieve their unique minima in $[0,1]$ at $\hat{p}$, $p_1$ and $p_2$, respectively. Furthermore, the function $g : [0,1] \to (0,\infty)$ is strictly convex, and has a unique minimizer in $[0,1]$ at $p_0 \in (0,1)$.*

*Proof.* We start with the functions $\log b_j(p)$. By routine calculus, the first and second derivatives are

$$\begin{aligned}
\frac{\partial}{\partial p}\log b_0(p) &= \log\left(\frac{p(1-p_1)(1-p_2)}{(1-p)p_1p_2}\right), \\
\frac{\partial}{\partial p}\log b_i(p) &= \log\left(\frac{p(1-p_i)}{(1-p)p_i}\right) && \text{for } i\in\{1,2\}, \\
\frac{\partial^2}{\partial^2 p}\log b_j(p) &= \frac{1}{p(1-p)} && \text{for } j\in\{0,1,2\}.
\end{aligned} \tag{3.94}$$

Note that the $\log b_j(p)$ with $j\in\{0,1,2\}$ are all strictly convex functions for $p\in[0,1]$, since each function is continuous on $[0,1]$ with a strictly positive second derivative for $p\in(0,1)$. Note that the first derivatives (3.94) of the $\log b_j(p)$ are all negative near 0 and positive near 1. So by solving for the zeroes of the first derivative, it follows that $\log b_0(p)$, $\log b_1(p)$ and $\log b_2(p)$ achieve their unique minima at $\hat{p}$, $p_1$ and $p_2$, respectively.

We now turn to the function $g(p)$. Letting $f_0(x)=\log b_0(x)$, $f_1(x)=2\log b_1(x)$ and $f_2(x)=2\log b_2(x)$, note that for any $a,b\in[0,1]$ and $t\in(0,1)$ with $a\neq b$, by strict convexity of the functions $f_j(x)$ it follows that

$$\begin{aligned}
g(ta+(1-t)b) &= \max_{j\in\{0,1,2\}}\{f_j(ta+(1-t)b)\} \\
&< \max_{j\in\{0,1,2\}}\{tf_j(a)+(1-t)f_j(b)\} \\
&\leq tg(a)+(1-t)g(b),
\end{aligned}$$

which establishes that $g(p)$ is a strictly convex function for $p\in[0,1]$. As uniqueness of the minimizer $p_0$ of $g(p)$ over $p\in[0,1]$ follows from strict convexity, it suffices to check that the minimum of $g(p)$ is not attained at the endpoints $p\in\{0,1\}$. This follows from the behavior of the first derivatives (3.94) of the $\log b_j(p)$ established above, which imply that $g(p)$ is decreasing near 0 and increasing near 1. Finally, it remains to determine the range of $g(p)$ for all $p\in[0,1]$: the upper bound $g(p)\leq\max\{g(0),g(1)\}<\infty$ follows from strict convexity of $g(p)$ and the convention $0\log 0=0$ (mentioned at the beginning of Section 3.4), and the lower bound $g(p)\geq\log b_0(p)\geq\log b_0(\hat{p})>0$ follows from the properties of $\log b_0(p)$ established above. □

*Proof of Lemma 3.3.1.* **(i):** From Lemma 3.4.2 we know that $g : [0,1] \to (0,\infty)$ has a unique minimizer $p_0 \in (0,1)$, and that $g(p)$ is strictly convex, which together readily establishes Lemma 3.3.1 (i).

**(ii):** From Lemma 3.4.2, for any $p \in [0,1]$ we know that $b_0(p) \geq b_0(\hat{p}) > 1$ and $b_i(p) \geq b_i(p_i) \geq 1$ for $i \in \{1,2\}$. This enables us to pick $\eta = \eta(p_1,p_2) > 0$ sufficiently small such that, for all sufficiently large $N$, we have

$$(4\log_{b_0(\hat{p})} N + 1) \cdot (1+\eta)^{4\log_{b_0(\hat{p})} N} \leq N \qquad \text{and} \qquad \log b_0(\hat{p}) \geq 3\log(1+\eta). \tag{3.95}$$

Setting $\lambda := 1+\eta$, we first analyze properties of $x_N^{(0)}(p)$. Since $\log b_0(p) \geq \log b_0(\hat{p}) > \log\lambda > 0$, by inspecting the definition (3.10) of $x_N^{(0)}(p)$ we infer that, for all sufficiently large $N$ (depending only on $\lambda$), we have

$$x_N^{(0)}(p) \leq 4\log_{b_0(\hat{p})} N + 1, \tag{3.96}$$

$$x_N^{(0)}(p) = 4\log_{b_0(p)} N + O(\log\log N), \tag{3.97}$$

where the asymptotic estimate in (3.97) is uniform in $p$, i.e., the implicit error term $O(\log\log N)$ depends only on $p_1, p_2$ (and not on $p$).

We next turn to $x_N^{(i)}(p)$ with $i \in \{1,2\}$, where $b_i = b_i(p) \geq 1$ holds. Using the definition (3.12) it is easy to see that $x_N^{(i)}(p) > 1$ holds for all $N \geq 1$ (since $x_N^{(i)}(p) \leq 1$ implies that $e \leq eN \leq 1 \cdot b_i^{-(1-x_N^{(i)}(p))/2} \leq 1$). For technical reasons we now distinguish whether $b_i = b_i(p)$ is smaller or larger than $\lambda$. We start with the case $1 \leq b_i < \lambda$, where using (3.12) and (3.95) we infer that

$$x_N^{(i)}(p) \cdot \lambda^{(x_N^{(i)}(p)-1)/2} \geq eN > (4\log_{b_0(\hat{p})} N + 1) \cdot \lambda^{[(4\log_{b_0(\hat{p})} N+1)-1]/2},$$

which in turn, by noting that $x\lambda^{(x-1)/2}$ is increasing for $x \geq 1$, implies together with (3.96) that

$$x_N^{(i)}(p) > 4\log_{b_0(\hat{p})} N + 1 \geq x_N^{(0)}(p). \tag{3.98}$$

In this case, gearing up towards (3.50), using (3.95) we also infer that

$$\frac{2\log N}{\log b_i(p)} \geq \frac{2\log N}{\log \lambda} > \frac{4\log N}{\log b_0(p)}. \tag{3.99}$$

We next consider the case $b_i \geq \lambda > 1$. Applying a bootstrapping argument to the implicit definition (3.12) of $x_N^{(i)}(p)$, i.e., $eN = x_N^{(i)}(p) b_i^{(x_N^{(i)}(p)-1)/2}$, for all sufficiently large $N$ (depending only on $\lambda$) it follows that

$$\begin{aligned}
x_N^{(i)}(p) &= 2\log_{b_i} N - 2\log_{b_i} x_N^{(i)}(p) + 2\log_{b_i} e + 1 \\
&= 2\log_{b_i} N - 2\log_{b_i}(2\log_{b_i} N - 2\log_{b_i} x_N^{(i)}(p) + 2\log_{b_i} e + 1) + 2\log_{b_i}(e) + 1 \\
&= 2\log_{b_i} N - 2\log_{b_i}\log_{b_i} N - 2\log_{b_i}(2/e) + 1 + O\left(\frac{\log\log N}{\log N}\right) \\
&= 2\log_{b_i} N + O(\log\log N),
\end{aligned} \tag{3.100}$$

where all the asymptotic estimates in (3.100) are uniform, i.e., do not depend on $p$ (here we exploit that $b_i \geq \lambda > 1$ in this case). In both cases, by combining (3.97) with (3.98)–(3.99) and (3.100), we obtain the uniform estimate

$$\min\{x_N^{(0)}(p), x_N^{(i)}(p)\} = \min\left\{\frac{4\log N}{\log b_0(p)}, \frac{2\log N}{\log b_i(p)}\right\} + O(\log\log N). \tag{3.101}$$

Applying (3.101) for both $i \in \{1,2\}$ then establishes the desired estimate (3.50) by definition of $g(p)$, completing the proof of Lemma 3.3.1 (ii). □

*Proof of Remark 3.1.5.* The definition of $n_N$ is exactly as in the proof of Theorem 3.1.3, so we only need to establish the asymptotic estimate (3.13). Here the crux is that from (3.98) we know that $x_N^{(i)}(p) \leq x_N^{(0)}(p)$ implies $b_i \geq \lambda$, so that estimate (3.100) applies to $x_N^{(i)}(p)$, which immediately establishes (3.13), as desired. □

### 3.4.2 Proofs of Corollary 3.1.4 and Lemma 3.4.1

*Proof of Corollary 3.1.4.* By assumption, Lemma 3.4.1 (i) implies that $n_N = x_N^{(0)}(\hat{p})$ for all $N$ large enough, Invoking Theorem 3.1.3, it thus only remains to verify that

$$\log b_0(\hat{p}) > \max\{2\log b_1(\hat{p}), 2\log b_2(\hat{p})\}$$

holds when $p_1 = p_2 = p$. To this end, we start with the trivial inequality

$$\frac{p^4 + (1-p)^4}{p^2 + (1-p)^2} < p^2 + (1-p)^2.$$

Since $\log(x)$ is a concave function, using $\hat{p} = \hat{p}(p,p) = p^2/(p^2 + (1-p)^2)$ and Jensen's inequality it follows that

$$\hat{p}\log p^2 + (1-\hat{p})\log(1-p)^2 \le \log\left(\frac{p^4 + (1-p)^4}{p^2 + (1-p)^2}\right) < \log\bigl(p^2 + (1-p)^2\bigr).$$

Exponentiating both sides, we infer that

$$p^{2\hat{p}}(1-p)^{2(1-\hat{p})} < p^2 + (1-p)^2.$$

Dividing both sides by $(p^2 + (1-p)^2)^2 > 0$, using $p_1 = p_2 = p$ we conclude that

$$b_1(\hat{p})^2 = b_2(\hat{p})^2 = \left(\frac{\hat{p}}{p}\right)^{2\hat{p}} \left(\frac{1-\hat{p}}{1-p}\right)^{2(1-\hat{p})} = \frac{p^{2\hat{p}}(1-p)^{2(1-\hat{p})}}{(p^2 + (1-p)^2)^2} < \frac{1}{p^2 + (1-p)^2} = b_0(\hat{p}),$$

which establishes that $\log b_0(\hat{p}) > \max\{2\log b_1(\hat{p}), 2\log b_2(\hat{p})\}$, as desired. $\square$

*Proof of Lemma 3.4.1.* (i): From Lemma 3.4.2, we know that $\log b_0(p)$ achieves its unique minimum at $p = \hat{p}$. Our assumption implies that $g(\hat{p}) = \log b_0(\hat{p})$. For any $p \in [0,1]$ we thus have

$$g(p) \ge \log b_0(p) \ge \log b_0(\hat{p}) = g(\hat{p}),$$

which establishes that $\hat{p}$ is the unique (see Lemma 3.4.2) minimizer $p_0$ of the function $g(p)$.

It remains to prove that $n_N = x_N^{(0)}(\hat{p})$ for all large enough $N$. Our assumption implies $\frac{1}{4}\log b_0(\hat{p}) > \max\{\frac{1}{2}\log b_1(\hat{p}), \frac{1}{2}\log b_2(\hat{p})\}$, which by comparing the first order terms in (3.101) implies

$$x_N^{(0)}(\hat{p}) < \min\left\{x_N^{(1)}(\hat{p}), x_N^{(2)}(\hat{p})\right\} \tag{3.102}$$

for all $N$ large enough (depending only on the functions $\log b_j(\hat{p})$ and thus $p_1, p_2$). Fix $p \in [0,1]$ with $p \neq \hat{p}$. Setting $r := \log b_0(\hat{p}) > 0$ and $\gamma := \log b_0(p) - r > 0$ (where $\gamma > 0$ follows from $p \neq \hat{p}$, as $\hat{p}$ is the unique minimizer of $\log b_0(p)$ by Lemma 3.4.2), using $\log(1 + \gamma/r) \le \gamma/r$ it follows for all $N$ large enough (depending only on $\log b_0(\hat{p})$ and thus $p_1, p_2$) that

$$
\begin{aligned}
&x_N^{(0)}(\hat{p}) - x_N^{(0)}(p) \\
&= \frac{4\log N - 2\log\log N - 2\log(4/e) + 2\log r}{r} - \frac{4\log N - 2\log\log N - 2\log(4/e) + 2\log(r+\gamma)}{r+\gamma} \\
&= \frac{\gamma}{r(r+\gamma)}\left[4\log N - 2\log\log N - 2\log(4/e) - \frac{2r}{\gamma}\log\left(1+\frac{\gamma}{r}\right) + 2\log r\right] \\
&\geq \frac{\gamma}{r(r+\gamma)}\left[2\log N - 1 + 2\log\log b_0(\hat{p})\right] \geq \frac{\gamma \log N}{r(r+\gamma)} > 0.
\end{aligned}
\tag{3.103}
$$

Combining inequalities (3.102) and (3.103) with the definition (3.11) of $n_N$, it then follows that

$$
x_N^{(0)}(\hat{p}) = \min\left\{x_N^{(0)}(\hat{p}),\, x_N^{(1)}(\hat{p}),\, x_N^{(2)}(\hat{p})\right\} \leq n_N \leq \max_{p\in[0,1]} x_N^{(0)}(p) = x_N^{(0)}(\hat{p}).
$$

Hence $n_N = x_N^{(0)}(\hat{p})$ for all large enough $N$ (depending only on $p_1, p_2$), completing the proof of Lemma 3.4.1 (i).

(ii): First, we show that the solution $p^*$ of $\log b_0(p) = 2\log b_i(p)$ is unique, and that $p^*$ lies between $\hat{p}$ and $p_i$, with $p^* \neq p_i$. To this end we introduce the auxiliary function

$$
h(p) := 2\log b_1(p) - \log b_0(p) = p\log\left(\frac{pp_1p_2}{p_i^2}\right) + (1-p)\log\left(\frac{(1-p)(1-p_1)(1-p_2)}{(1-p_i)^2}\right).
$$

Using (3.94) we see that the second derivative is $h''(p) = 1/[p(1-p)]$. Hence $h(p)$ is a strictly convex function for $p \in [0,1]$, since it is a continuous function on $[0,1]$ with a strictly positive second derivative for $p \in (0,1)$. The above proof of Corollary 3.1.4 shows that $p_1 = p_2$ falls in the previous case (i), so that we here have $p_1 \neq p_2$. Consequently, $h(0)$ and $h(1)$ are both not equal to 0, and in particular have different signs (using $0\log 0 = 0$, as mentioned at the beginning of Section 3.4). Using strict convexity it follows that $h(p)$ has a unique zero in $[0,1]$, which by construction is the unique solution $p^*$ of $\log b_0(p) = 2\log b_i(p)$. Furthermore, since our assumptions imply $2\log b_i(\hat{p}) \geq \log b_0(\hat{p})$ and $p_1, p_2 \in (0,1)$, for $j \in \{1,2\}\setminus\{i\}$ it follows that

$$
h(\hat{p}) \geq 0 \qquad \text{and} \qquad h(p_i) = p_i \log p_j + (1-p_i)\log(1-p_j) < 0.
$$

which establishes that $p^*$ lies between $\hat{p}$ and $p_i$, with $p^* \neq p_i$.

Second, we claim that $f(p) := \max\{\log b_0(p), 2\log b_i(p)\}$ is decreasing in $[0, p^*]$ and increasing in $[p^*, 1]$, which together with $\log b_0(p^*) = 2\log b_i(p^*)$ establishes the useful fact

$$\min_{p\in[0,1]} \max\big\{\log b_0(p),\, 2\log b_i(p)\big\} = 2\log b_i(p^*). \tag{3.104}$$

By Lemma 3.4.2, $\log b_0(p)$ is decreasing in $[0, \hat{p}]$ and increasing in $[\hat{p}, 1]$. Furthermore, $2\log b_i(p)$ is decreasing in $[0, p_1]$ and increasing in $[p_1, 1]$. Hence $f(p)$ is increasing in $[\max\{p_1, \hat{p}\}, 1]$. Furthermore, recalling that $p^*$ lies between $\hat{p}$ and $p_1$, it follows that in $[p^*, \max\{p_1, \hat{p}\}]$ one of $\log b_0(p)$ or $2\log b_i(p)$ is increasing, whereas the other one is decreasing. Since $\log b_0(p^*) = 2\log b_i(p^*)$, it follows that $f(p)$ is increasing in $[p^*, \max\{p_1, \hat{p}\}]$, establishing that $f(p)$ is increasing in $[p^*, 1]$. A similar argument shows that $f(p)$ is decreasing in $[0, p^*]$.

Third, we show that $p^*$ is the minimizer of $g(p)$. To see this, note that $\log b_0(p^*) = 2\log b_i(p^*)$ implies

$$b_0(p^*) = b_0(p^*) \cdot \frac{b_0(p^*)}{b_i(p^*)^2} = \left[\left(\frac{1}{p_j}\right)^p \left(\frac{1}{1-p_j}\right)^{1-p}\right]^2 \geq b_j(p^*)^2$$

for $j \in \{1,2\} \setminus \{i\}$, so that $g(p^*) = f(p^*) = 2\log b_i(p^*)$. Using (3.104), for any $p \in [0,1]$ it then follows that

$$g(p) \geq \max\big\{\log b_0(p),\, 2\log b_i(p)\big\} \geq 2\log b_i(p^*) = g(p^*),$$

which establishes that $p^*$ is the unique (see Lemma 3.4.2) minimizer $p_0$ of the function $g(p)$.

Finally, from Lemma 3.3.1(ii) and (3.51), using that $p_0 = p^* \in (0,1)$ and $g(p_0) = 2\log b_i(p_0) > 0$, it follows that

$$n_N = \frac{4\log N}{g(p_0)} + O(\log\log N) = x_N^{(i)}(p_0) + O(\log\log N) = (1+o(1))2\log_{b_i(p_0)} N,$$

which establishes (3.93) and thus completes the proof of Lemma 3.4.1(ii). □

## 3.5 Proof of Lemma 3.1.6: Pseudorandom properties

*Proof of Lemma 3.1.6.* We start with the binomial random $G_{n,p}$, whose number of edges has a binomial distribution with expected value $\binom{n}{2} \cdot p$. Using Chebyshev's inequality (or Chernoff bounds), it easily follows that

$$\mathbb{P}(G_{n,p} \notin \mathscr{F}_{n,p}) = o(1). \tag{3.105}$$

Furthermore, the estimate

$$\mathbb{P}(G_{n,p} \text{ is not asymmetric}) \leq e^{-\Theta(np(1-p))} \tag{3.106}$$

follows, for instance, by the proof of Theorem 3.1 in [KSV02] when replacing the $\varepsilon$ in their proof by some small constant (their proof in fact works for any $p = p(n)$ satisfying $np(1-p) \gg \log n$). Note that the induced subgraph $G_{n,p}[L]$ has the same distribution as $G_{|L|,p}$. Taking a union bound over all possible vertex-subsets $L \subseteq [n]$ of size $\ell := |L| \geq n - n^{2/3}$, using (3.106) and $\binom{n}{\ell} = \binom{n}{n-\ell} \leq n^{n-\ell} \leq n^{n^{2/3}}$ we obtain

$$\mathbb{P}(G_{n,p} \notin \mathscr{A}_n) \leq \sum_{n-n^{2/3} \leq \ell \leq n} \binom{n}{\ell} \cdot e^{-\Theta(\ell p(1-p))} \leq e^{-\Theta(np(1-p))} = o(1), \tag{3.107}$$

which in fact holds for any edge-probability $p = p(n)$ satisfying $n^{1/3}p(1-p) \gg \log n$. Combining (3.105) and (3.107) establishes that $G_{n,p} \in \mathscr{A}_n \cap \mathscr{F}_{n,p}$ whp.

We next consider the uniform random graph $G_{n,m}$. Using Pittel's inequality (see [Bol01, Theorem 2.2]) with $p' = p'(n) := m/\binom{n}{2} \in [\gamma, 1-\gamma]$, it routinely follows from inequality (3.107) that

$$\mathbb{P}(G_{n,m} \notin \mathscr{A}_n) \leq 3\sqrt{m} \cdot \mathbb{P}(G_{n,p'} \notin \mathscr{A}_n) \leq O(n) \cdot e^{-\Theta(np'(1-p'))} = o(1). \tag{3.108}$$

Turning to $\mathscr{E}_{n,m}$, fix any $L \subseteq [n]$ and write $\ell := |L|$ for the size of $L$, as before. To estimate the number of edges contained in the induced subgraph $G_{n,m}[L]$, we use that by definition of $G_{n,m}$ we have

$$\underbrace{|E(G_{n,m})|}_{=m} = |E(G_{n,m}[L])| + \underbrace{|E(G_{n,m}) \setminus E(G_{n,m}[L])|}_{=:X_L}. \tag{3.109}$$

For $L=[n]$ this already gives $|E(G_{n,m}[L])| = \binom{|L|}{2}m/\binom{n}{2}$, so it remains to consider the case $L \subsetneq [n]$. The random variable $X_L$ has a hypergeometric distribution with parameters $K := \binom{n}{2}$, $m = \Theta(n^2)$ and

$$k_\ell := \binom{n}{2} - \binom{\ell}{2} = \frac{(n-\ell)(n+\ell-1)}{2},$$

with expected value $\mathbb{E}X_L = k_\ell \cdot m/K = \Theta(n(n-\ell))$, since we have $k_\ell$ random draws (without replacement) out of a total of $K = \binom{n}{2}$ potential edges, where $m \in [\gamma\binom{n}{2}, (1-\gamma)\binom{n}{2}]$ out of the potential edges are present. Since standard Chernoff bounds for binomial random variables with expected value $k_\ell \cdot m/K$ also apply to $X_L$, see [JŁR00, Theorem 2.1 and 2.10], it routinely follows that

$$\mathbb{P}\big(|X_L - k_\ell m/K| \geq n^{2/3}(n-\ell)\big) \leq 2\exp\left(\frac{-n^{4/3}(n-\ell)^2}{3\mathbb{E}X_L}\right) \leq e^{-\Theta(n^{1/3}(n-\ell))}.$$

Writing $\mathscr{E}'_{n,m}$ for the event that $|X_L - k_\ell m/K| < n^{2/3}(n-\ell)$ for all $L \subsetneq [n]$, using a standard union bound argument and $\binom{n}{\ell} = \binom{n}{n-\ell} \leq n^{n-\ell}$ it readily follows that

$$\mathbb{P}(G_{n,m} \notin \mathscr{E}'_{n,m}) \leq \sum_{0 \leq \ell \leq n-1} \binom{n}{\ell} \cdot e^{-\Theta(n^{1/3}(n-\ell))} = o(1). \tag{3.110}$$

Combining (3.108) and (3.110) establishes that $G_{n,m} \in \mathscr{A}_n \cap \mathscr{E}'_{n,m}$ whp. It remains to show that $G_{n,m} \in \mathscr{E}'_{n,m}$ implies $G_{n,m} \in \mathscr{E}_{n,m}$. To see this note that, for any $L \subsetneq [n]$, equation (3.109) and $G_{n,m} \in \mathscr{E}'_{n,m}$ imply

$$|E(G_{n,m}[L])| = m - X_L = m - \frac{\left[\binom{n}{2} - \binom{|L|}{2}\right]m}{\binom{n}{2}} \pm n^{2/3}(n-|L|) = \frac{\binom{|L|}{2}m}{\binom{n}{2}} \pm n^{2/3}(n-|L|),$$

which completes the proof of Lemma 3.1.6 (since this edge-estimate is trivially true for $L=[n]$, as discussed). □

## 3.6 Concluding remarks

In this paper we resolved the induced subgraph isomorphism problem and the maximum common induced subgraph problem for dense random graphs, i.e., with constant edge-probabilities. Besides the convergence questions already mentioned in Sections 3.2.2, 3.2.3

and 3.3, there are two main directions for further research: extensions of our main results to sparse random graphs, where edge-probabilities $p_i \to 0$ are allowed (Problem 3.6.1), and generalizations of the graph-sizes (Problem 3.6.2).

**Problem 3.6.1** (Edge-Sparsity)**.** *Prove versions of the induced subgraph isomorphism results Theorems 3.1.1–3.1.2 and maximum common induced subgraph result Theorem 3.1.3 for sparse random graphs.*

As a first step towards such sparse extensions of our main results, one can initially aim at slightly weaker results in the sparse case. For example, in Theorem 3.1.3 one could instead try to show that typically $I_N = (1+o(1))\Lambda_N$ for some explicit $\Lambda_N = \Lambda_N(p_1, p_2)$. Furthermore, in Theorem 3.1.1 one could instead try to determine some explicit $n^* = n^*(N, p_1, p_2)$ such that $\mathbb{P}(G_{n,p_1} \sqsubseteq G_{N,p_2})$ changes from $1-o(1)$ to $o(1)$ for $n \le (1-\varepsilon)n^*$ and $n \ge (1+\varepsilon)n^*$, respectively. As another example, in Theorem 3.1.3 with $p_1 = p_2 = p$ and $p = p(N) \to 0$ we wonder if two-point concentration of $I_N$ remains valid for $p \gg N^{-2/3+\varepsilon}$.

**Problem 3.6.2** (Generalization)**.** *Fix constants $p_1, p_2 \in (0,1)$. Determine, as $N_1, N_2 \to \infty$, the typical size $I_{N_1,N_2}$ of the maximum common induced subgraph of the independent random graphs $G_{N_1,p_1}$ and $G_{N_2,p_2}$.*

This problem aims at a common generalization of our main results, since for $N_1 \le N_2$ the two extreme cases $N_1 \le 2\log_a N_2 - \omega(1)$ and $N_1 = N_2$ are already covered by Theorems 3.1.1 and 3.1.3. So the real question is what happens in-between: will, similar as in this paper, new concentration phenomena occur?

### 3.6.1 Acknowledgements

Chapter 3 is a reprint, in full, of the material as it appears in the preprint *Isomorphisms between random graphs*. This preprint was co-authored by the dissertation author together with Erlang Surya and Lutz Warnke. It was previously submitted for publication.

# Bibliography


[AGLM14] Maria Axenovich, András Gyárfás, Hong Liu, and Dhruv Mubayi. Multicolor Ramsey numbers for triple systems. *Discrete Math.*, 322:69–77, 2014.

[AKS80] Miklós Ajtai, János Komlós, and Endre Szemerédi. A note on Ramsey numbers. *J. Combin. Theory Ser. A*, 29(3):354–360, 1980.

[Alo17] Noga Alon. Asymptotically optimal induced universal graphs. *Geom. Funct. Anal.*, 27(1):1–32, 2017.

[Alo23] Noga Alon, 2023. Personal communication.

[AS04] Noga Alon and Asaf Shapira. Testing subgraphs in directed graphs. *J. Comput. System Sci.*, 69(3):353–382, 2004.

[BBSV19] Paul Balister, Béla Bollobás, Julian Sahasrabudhe, and Alexander Veremyev. Dense subgraphs in random graphs. *Discrete Appl. Math.*, 260:66–74, 2019.

[BGP$^+$13] Vincenzo Bonnici, Rosalba Giugno, Alfredo Pulvirenti, Dennis Shasha, and Alfredo Ferro. A subgraph isomorphism algorithm and its application to biochemical data. *BMC Bioinform.*, 14:1–13, 2013.

[BK10] Tom Bohman and Peter Keevash. The early evolution of the $H$-free process. *Invent. Math.*, 181(2):291–336, 2010.

[Boh09] Tom Bohman. The triangle-free process. *Adv. Math.*, 221(5):1653–1677, 2009.

[Bol01] Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[CD23] Sourav Chatterjee and Persi Diaconis. Isomorphisms between random graphs. *J. Combin. Theory Ser. B*, 160:144–162, 2023.

[CFS10] David Conlon, Jacob Fox, and Benny Sudakov. Hypergraph Ramsey numbers. *J. Amer. Math. Soc.*, 23(1):247–266, 2010.

[CFSV04] Donatello Conte, Pasquale Foggia, Carlo Sansone, and Mario Vento. Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.*, 18:265–298, 2004.

[CG98] Fan Chung and Ron Graham. *Erdős on graphs: his legacy of unsolved problems*. A K Peters, Ltd., Wellesley, MA, 1998.

[CH78] Louis Caccetta and R. Häggkvist. On minimal digraphs with given girth. In *Proceedings of the Ninth Southeastern Conference on Combinatorics, Graph Theory, and Computing (Florida Atlantic Univ., Boca Raton, Fla., 1978)*, volume XXI of *Congress. Numer.*, pages 181–187. Utilitas Math., Winnipeg, MB, 1978.

[CH99] Diane Cook and Lawrence Holder. Substructure discovery using minimum description length and background knowledge. *J. Artif. Intell. Res.*, 1:231–255, 1999.

[Chu97] F. R. K. Chung. Open problems of Paul Erdős in graph theory. *J. Graph Theory*, 25(1):3–36, 1997.

[Con09] David Conlon. A new upper bound for diagonal Ramsey numbers. *Ann. of Math. (2)*, 170(2):941–960, 2009.

[DJR17] E. Davies, M. Jenssen, and B. Roberts. Multicolour Ramsey numbers of paths and even cycles. *European J. Combin.*, 63:124–133, 2017.

[DSdlH+11] Guillaume Damiand, Christine Solnon, Colin de la Higuera, Jean-Christophe Janodet, and Émilie Samuel. Polynomial algorithms for subisomorphism of nd open combinatorial maps. *Comput. Vis. Image Underst.*, 115:996–1010, 2011.

[EH63] P. Erdős and H. Hanani. On a limit theorem in combinatorial analysis. *Publ. Math. Debrecen*, 10:10–13, 1963.

[EH72] P. Erdős and A. Hajnal. On Ramsey like theorems. Problems and results. In *Combinatorics (Proc. Conf. Combinatorial Math., Math. Inst., Oxford, 1972)*, pages 123–140. Inst. Math. Appl., Southend-on-Sea, 1972.

[EHR65] P. Erdős, A. Hajnal, and R. Rado. Partition relations for cardinal numbers. *Acta Math. Acad. Sci. Hungar.*, 16:93–196, 1965.

[EKR61] P. Erdős, Chao Ko, and R. Rado. Intersection theorems for systems of finite sets. *Quart. J. Math. Oxford Ser. (2)*, 12:313–320, 1961.

[ER52] P. Erdős and R. Rado. Combinatorial theorems on classifications of subsets of a given set. *Proc. London Math. Soc. (3)*, 2:417–439, 1952.

[ER63] P. Erdős and A. Rényi. Asymmetric graphs. *Acta Math. Acad. Sci. Hungar.*, 14:295–315, 1963.

[ER11] Christian Ehrlich and Matthias Rarey. Maximum common subgraph isomorphism algorithms and their applications in molecular science: A review. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, 1:68–79, 2011.

[Erd47] P. Erdős. Some remarks on the theory of graphs. *Bull. Amer. Math. Soc.*, 53:292–294, 1947.

[ES35] P. Erdős and G. Szekeres. A combinatorial problem in geometry. *Compositio Math.*, 2:463–470, 1935.

[FF83] Peter Frankl and Zoltán Füredi. A new generalization of the Erdős-Ko-Rado theorem. *Combinatorica*, 3(3-4):341–349, 1983.

[FJK+21] Zoltán Füredi, Tao Jiang, Alexandr Kostochka, Dhruv Mubayi, and Jacques Verstraëte. Extremal problems for convex geometric hypergraphs and ordered hypergraphs. *Canad. J. Math.*, 73(6):1648–1666, 2021.

[FKM14] Nikolaos Fountoulakis, Ross J. Kang, and Colin McDiarmid. Largest sparse subgraphs of random graphs. *European J. Combin.*, 35:232–244, 2014.

[FO11] Zoltán Füredi and Lale Özkahya. Unavoidable subhypergraphs: **a**-clusters. *J. Combin. Theory Ser. A*, 118(8):2246–2256, 2011.

[GBB+13] Rosalba Giugno, Vincenzo Bonnici, Nicola Bombieri, Alfredo Pulvirenti, Alfredo Ferro, and Dennis Shasha. GRAPES: A software for parallel searching on biological graphs targeting multi-core architectures. *PLoS One*, 8(10):1–11, 2013.

[GR12] Andras Gyárfás and Ghaffar Raeisi. The Ramsey number of loose triangles and quadrangles in hypergraphs. *Electron. J. Combin.*, 19(2):Paper 30, 9, 2012.

[HKN17] Jan Hladký, Daniel Král, and Sergey Norin. Counting flags in triangle-free digraphs. *Combinatorica*, 37(1):49–76, 2017.

[Irv74] Robert W. Irving. Generalised Ramsey numbers for small graphs. *Discrete Math.*, 9:251–264, 1974.

[Jac15] Eliza Jackowska. The 3-color Ramsey number for a 3-uniform loose path of length 3. *Australas. J. Combin.*, 63:314–320, 2015.

[Jan22] Svante Janson, 2022. Personal communication.

[JŁR00] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.

[JPR16] Eliza Jackowska, Joanna Polcyn, and Andrzej Ruciński. Turán numbers for 3-uniform linear paths of length 3. *Electron. J. Combin.*, 23(2):Paper 2.30, 18, 2016.

[JPR17] Eliza Jackowska, Joanna Polcyn, and Andrzej Ruciński. Multicolor Ramsey numbers and restricted Turán numbers for the loose 3-uniform path of length three. *Electron. J. Combin.*, 24(3):Paper No. 3.5, 21, 2017.

[KS19] Charlotte Knierim and Pascal Su. Improved bounds on the multicolor Ramsey numbers of paths and even cycles. *Electron. J. Combin.*, 26(1):Paper No. 1.26, 17, 2019.

[KSV02] Jeong Han Kim, Benny Sudakov, and Van H. Vu. On the asymmetry of random regular graphs and random graphs. *Random Structures Algorithms*, 21:216–224, 2002. Random structures and algorithms (Poznan, 2001).

[ŁP17] Tomasz Łuczak and Joanna Polcyn. On the multicolor Ramsey number for 3-paths of length three. *Electron. J. Combin.*, 24(1):Paper No. 1.27, 4, 2017.

[ŁP18] Tomasz Łuczak and Joanna Polcyn. The multipartite Ramsey number for the 3-path of length three. *Discrete Math.*, 341(5):1270–1274, 2018.

[ŁP19] Tomasz Łuczak and Joanna Polcyn. Paths in hypergraphs: a rescaling phenomenon. *SIAM J. Discrete Math.*, 33(4):2251–2266, 2019.

[MPST18] Ciaran McCreesh, Patrick Prosser, Christine Solnon, and James Trimble. When subgraph isomorphism is really hard, and why this matters for graph databases. *J. Artificial Intelligence Res.*, 61:723–759, 2018.

[MPT16] Ciaran McCreesh, Patrick Prosser, and James Trimble. Heuristics and really hard instances for subgraph isomorphism problems. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 631–638. AAAI Press, 2016.

[MS18] Dhruv Mubayi and Andrew Suk. New lower bounds for hypergraph Ramsey numbers. *Bull. Lond. Math. Soc.*, 50(2):189–201, 2018.

[MS20] Dhruv Mubayi and Andrew Suk. The Erdős–Hajnal hypergraph Ramsey problem. *J. Eur. Math. Soc. (JEMS)*, 22(4):1247–1259, 2020.

[Mub06] Dhruv Mubayi. Erdős–Ko–Rado for three sets. *J. Combin. Theory Ser. A*, 113(3):547–550, 2006.

[MV05] Dhruv Mubayi and Jacques Verstraëte. Proof of a conjecture of Erdős on triangles in set-systems. *Combinatorica*, 25(5):599–614, 2005.

[Pol17] Joanna Polcyn. One more Turán number and Ramsey number for the loose 3-uniform path of length three. *Discuss. Math. Graph Theory*, 37(2):443–464, 2017.

[PR17a] Joanna Polcyn and Andrzej Ruciński. A hierarchy of maximal intersecting triple systems. *Opuscula Math.*, 37(4):597–608, 2017.

[PR17b] Joanna Polcyn and Andrzej Ruciński. Refined Turán numbers and Ramsey numbers for the loose 3-uniform path of length three. *Discrete Math.*, 340(2):107–118, 2017.

[RW02] John Raymond and Peter Willett. Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *J. Comput.-Aided Mol. Des.*, 16:521–33, 2002.

[Sár16] Gábor N. Sárközy. On the multi-colored Ramsey numbers of paths and even cycles. *Electron. J. Combin.*, 23(3):Paper 3.53, 9, 2016.

[Spe78] Joel Spencer. Asymptotic lower bounds for Ramsey functions. *Discrete Math.*, 20(1):69–76, 1977/78.