

# A note on the random greedy independent set algorithm

Patrick Bennett\*

Tom Bohman†

## Abstract

Let  $r$  be a fixed constant and let  $\mathcal{H}$  be an  $r$ -uniform,  $D$ -regular hypergraph on  $N$  vertices. Assume further that  $D > N^\epsilon$  for some  $\epsilon > 0$ . Consider the random greedy algorithm for forming an independent set in  $\mathcal{H}$ . An independent set is chosen at random by iteratively choosing vertices at random to be in the independent set. At each step we chose a vertex uniformly at random from the collection of vertices that could be added to the independent set (i.e. the collection of vertices  $v$  with the property that  $v$  is not in the current independent set  $I$  and  $I \cup \{v\}$  contains no edge in  $\mathcal{H}$ ). Note that this process terminates at a maximal subset of vertices with the property that this set contains no edge of  $\mathcal{H}$ ; that is, the process terminates at a maximal independent set.

We prove that if  $\mathcal{H}$  satisfies certain degree and codegree conditions then there are  $\Omega\left(N \cdot ((\log N)/D)^{\frac{1}{r-1}}\right)$  vertices in the independent set produced by the random greedy algorithm with high probability. This result generalizes a lower bound on the number of steps in the  $H$ -free process due to Bohman and Keevash and produces objects of interest in additive combinatorics.

## 1 Introduction

Consider the random greedy algorithm for finding a maximal independent set in a hypergraph. Let  $\mathcal{H}$  be a hypergraph on vertex set  $V$ . (I.e.  $\mathcal{H}$  is a collection of subsets of  $V$ . The sets in this collection are the *edges* of  $\mathcal{H}$ ). An *independent set* in  $\mathcal{H}$  is a set  $I \subseteq V$  such that  $I$  contains no edge of  $\mathcal{H}$ . The random greedy algorithm forms a maximal independent set in  $\mathcal{H}$  by iteratively choosing vertices at random to be vertices in the independent set. To be precise, we begin with  $\mathcal{H}(0) = \mathcal{H}$ ,  $V(0) = V$  and  $I(0) = \emptyset$ . Given independent set  $I(i)$  and hypergraph  $\mathcal{H}(i)$  on vertex set  $V(i)$ , a vertex  $v \in V(i)$  is

---

\*Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: [ptbennet@andrew.cmu.edu](mailto:ptbennet@andrew.cmu.edu). Research supported in part by NSF grants DMS-1001638 and DMS-1362785.

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA. Email: [tbohman@math.cmu.edu](mailto:tbohman@math.cmu.edu). Research supported in part by NSF grant DMS-1001638.

chosen uniformly at random and added to  $I(i)$  to form  $I(i+1)$ . The vertex set  $V(i+1)$  is set equal to  $V(i)$  less  $v$  and every vertex  $u$  such that the pair  $\{u, v\}$  is an edge of  $\mathcal{H}(i)$ . Finally the hypergraph  $\mathcal{H}(i+1)$  is formed from  $\mathcal{H}(i)$  by

1. removing  $v$  from every edge in  $\mathcal{H}(i)$  that contains  $v$  and at least 2 other vertices, and
2. removing every edge that contains a vertex  $u$  such that the pair  $\{u, v\}$  is an edge of  $\mathcal{H}(i)$ .

The process terminates when  $V(i)$  is empty. At this point  $I(i)$  is a maximal independent set in  $\mathcal{H}$ .

A number of problems in combinatorics can be stated in terms of maximal independent sets in hypergraphs. In some of these situations, the random greedy algorithm produces such an independent set with desirable properties. For example, the best known lower bounds on the Turán numbers of some bipartite graphs as well as the best known lower bound on the off-diagonal graph Ramsey numbers  $R(s, t)$  (where  $s \geq 3$  is fixed and  $t$  is large) are given by objects produced by this algorithm. In these two cases the objects of interest are produced by an instance of the random greedy independent set algorithm known as the  $H$ -free process. Here we let  $H$  be a fixed 2-balanced graph (e.g.  $K_\ell$ ) and consider the hypergraph  $\mathcal{H}_H$  that has vertex set  $V = \binom{[n]}{2}$ , i.e. the edge set of the complete graph  $K_n$ , and edge set consisting of all copies of  $H$  in  $K_n$ . Note that in this context the random greedy independent set algorithm produces a graph on vertex set  $[n]$  (i.e. a subset of  $\binom{[n]}{2}$ ) that contains no copy of the graph  $H$ . Bohman and Keevash [6] gave an analysis of the  $H$ -free process for an arbitrary 2-balanced graph  $H$  that gives a lower bound on the number of steps in the process. In this note we extend that result to a more general setting. This generalization includes natural hypergraph variants of the  $H$ -free process as well as some processes that are of interest in number theory.

Following the intuition that guides the earlier work on the  $H$ -free process, our study of the random greedy independent set algorithm on a general  $D$ -regular,  $r$ -uniform hypergraph  $\mathcal{H}$  on vertex set  $V$  is guided by the following question:

To what extent does the independent set  $I(i)$  resemble a random subset  $S(i)$  of  $V$  chosen by simply taking  $\Pr(v \in S(i)) = i/N$ , independently, for all  $v \in V$ ?

Of course, if  $i$  is large enough then the set  $S(i)$  should contain many edges of  $\mathcal{H}$  while  $I(i)$  contains none. But are these sets similar with respect to other statistics? Consider, for example, the set of vertices  $V(i)$ , which is the set of vertices that remain eligible for inclusion in the independent set. A vertex  $w$  (that does not lie in  $I(i)$  itself) is in this set if there is no edge  $e \in \mathcal{H}$  such that  $w \in e$  and  $e \setminus \{w\} \subseteq I(i)$ . If  $I(i)$  resembles  $S(i)$  then the number of vertices that have this property should be roughly

$$|V| \left( 1 - \left( \frac{i}{N} \right)^{r-1} \right)^D \approx N \exp \left\{ -D \left( \frac{i}{N} \right)^{r-1} \right\}.$$

If this is indeed the case we would expect the algorithm to continue until

$$D \left( \frac{i}{N} \right)^{r-1} = \Omega(\log N).$$

Our main result is that if  $\mathcal{H}$  satisfies certain degree and codegree conditions this is indeed the case. And in the course of proving this result we establish a number of other similarities of  $I(i)$  and  $S(i)$ .

Define the *degree* of a set  $A \subset V$  to be the number of edges of  $\mathcal{H}$  that contain  $A$ . For  $a = 2, \dots, r-1$  we define  $\Delta_a(\mathcal{H})$  to be the maximum degree of  $A$  over  $A \in \binom{V}{a}$ . We also define the  $(r-1)$ -*codegree* of a pair of distinct vertices  $v, v'$  to be the number of edges  $e, e' \in \mathcal{H}$  such that  $v \in e \setminus e', v' \in e' \setminus e$  and  $|e \cap e'| = r-1$ . We let  $\Gamma(\mathcal{H})$  be the maximum  $(r-1)$ -codegree of  $\mathcal{H}$ .

**Theorem 1.1.** *Let  $r$  and  $\epsilon > 0$  be fixed. Let  $\mathcal{H}$  be a  $r$ -uniform,  $D$ -regular hypergraph on  $N$  vertices such that  $D > N^\epsilon$ . If*

$$\Delta_\ell(\mathcal{H}) < D^{\frac{r-\ell}{r-1}-\epsilon} \quad \text{for } \ell = 2, \dots, r-1 \quad (1)$$

*and  $\Gamma(\mathcal{H}) < D^{1-\epsilon}$  then the random greedy independent set algorithm produces an independent set  $I$  in  $\mathcal{H}$  with*

$$|I| = \Omega \left( N \cdot \left( \frac{\log N}{D} \right)^{\frac{1}{r-1}} \right) \quad (2)$$

*with probability  $1 - \exp \{-N^{\Omega(1)}\}$ .*

The proof of Theorem 1.1 is given in Section 4. Similar lower bounds on the independence numbers of  $r$ -uniform hypergraphs were established by Ajtai, Komlós, Pintz, Spencer and Szemerédi [1] and extended by Duke, Lefmann and Rödl [11]. However, these results are restricted to the case of simple hypergraphs (i.e.  $\Delta_2(\mathcal{H}) = 1$ ). We note that some condition beyond a vertex-degree condition is required to ensure existence of an independent set of size  $\omega(N/D^{\frac{1}{r-1}})$ ; for example, Cooper and Mubayi [9] recently gave a construction of a  $K_4^-$ -free 3-uniform hypergraph with independence number bounded above by  $2N/\Delta^{1/2}$ , where  $\Delta$  denotes maximum degree.

Consider the  $H$ -free process, where  $H$  is a graph with vertex set  $V_H$  and edge set  $E_H$ . Set  $v_H = |V_H|$  and  $e_H = |E_H|$ . Recall that  $H$  is strictly 2-balanced if and only if

$$\frac{e_H[W] - 1}{|W| - 2} < \frac{e_H - 1}{v_H - 2} \quad \text{for all } W \subsetneq V_H \text{ such that } |W| \geq 3, \quad (3)$$

where  $H[W]$  is the subgraph of  $H$  induced by  $W$ . For  $a \geq 2$  let  $v_a$  be the minimum number of vertices of  $H$  spanned by a set of  $a$  edges of  $H$ . We have

$$\Delta_a(\mathcal{H}_H) = \Theta(n^{v_H - v_a}) = \Theta \left( n^{(v_H - 2) \left[ 1 - \frac{v_a - 2}{v_H - 2} \right]} \right)$$

and

$$D^{\frac{e_H - a}{e_H - 1}} = \Theta \left( n^{\frac{(v_H - 2)(e_H - a)}{e_H - 1}} \right) = \Theta \left( n^{(v_H - 2) \left[ 1 - \frac{a - 1}{e_H - 1} \right]} \right).$$

Thus, we see that  $\mathcal{H}_H$  satisfies (1) if and only if  $(v_a - 2)/(v_H - 2) > (a - 1)/(e_H - 1)$  for all  $a \geq 2$ , which holds if and only if  $H$  is strictly 2-balanced. Thus Theorem 1.1 is a generalization of the lower bound on the number of steps in the  $H$ -free process for  $H$  strictly 2-balanced given by Bohman and Keevash [6].

We note that the hypotheses of Theorem 1.1 imply that  $\mathcal{H}$  is not too dense. Specifically, counting  $|\mathcal{H}|$  in two ways we have

$$\frac{ND}{r} = |\mathcal{H}| \leq \frac{1}{\binom{r}{2}} \binom{N}{2} D^{\frac{r-2}{r-1} - \epsilon}.$$

This implies that we have

$$N = \Omega \left( D^{\frac{1}{r-1} + \epsilon} \right). \quad (4)$$

We make use of this observation in the proofs below.

Some processes for which the degree and codegree conditions in Theorem 1.1 are relaxed have already been studied. A *diamond* is the graph obtained by removing an edge from  $K_4$ . The diamond-free process studied by Picollelli [21] is an example of an  $H$ -free process where the graph  $H$  is 2-balanced but not strictly 2-balanced. When  $H$  is a diamond then the hypergraph  $\mathcal{H}_H$  is 5-uniform and  $5(n - 2)(n - 3)/2$ -regular but has  $\Delta_3(\mathcal{H}_H) = 3(n - 3) = \Theta(D^{1/2})$ . For this process Picollelli [21] shows that the number of steps is larger than the bound given by (2) by a logarithmic factor. Bennett [2] has recent results on the sum-free process. This process is the random greedy independent set algorithm on the hypergraph which has vertex set  $\mathbb{Z}_n$  and edge set consisting of all solutions of the equations  $a + b = c$ . This hypergraph does not satisfy the codegree condition in Theorem 1.1. Since  $a + b = c$  implies  $(-a) + c = b$ , the 2-codegree of  $a$  and  $-a$  has the same order as the degree  $D$  of the hypergraph. Nevertheless, the lower bound (2) still holds for the sum-free process. In both of these processes, interesting irregularities in  $\mathcal{H}(i)$  (i.e. violations of our intuition that  $S(i)$  should resemble  $I(i)$ ) develop as the process evolves.

It is tempting to speculate that the lower bound in Theorem 1.1 gives the correct order of magnitude of the maximal independent set produced by the random greedy independent set algorithm for a broad class of hypergraphs  $\mathcal{H}$ . Bohman and Keevash conjecture that this is the case for the  $H$ -free process when  $H$  is strictly 2-balanced, but even this remains widely open. The conjecture has been verified in some special cases, including the  $K_3$ -free process [3], the  $K_4$ -free process [28, 30] and the  $C_\ell$ -free process for all  $\ell \geq 4$  [22, 23, 27].

In the interest of communicating a short and versatile proof, we make no attempt to optimize (or even explicitly state) the constant in the lower bound (2). For the same reason, we also refrain from considering hypergraphs that are only approximately  $D$ -regular. (It is clear that some loosening of the regularity condition in Theorem 1.1 is possible.) Our proof uses the so-called differential equations method for establishing

dynamic concentration and is a modest simplification of the earlier work of Bohman and Keevash. We do not establish self-correcting estimates, which are dynamic concentration inequalities with error bounds that improve as the underlying process evolves. Such estimates were first deployed by Telcs, Wormald and Zhou [25] (and, independently, in [8]). Bohman, Frieze and Lubetzky [4] developed a *critical interval* method for proving self-correcting estimates. Very recently, the critical interval method (and closely related methods) have been used to give a very detailed analysis of the triangle-removal process [5] and to determine the asymptotic number of edges in the  $K_3$ -free process [7, 12]. The latter gives an improvement on Kim's [19] celebrated lower bound on the Ramsey number  $R(3, t)$ .

Of course, the cardinality of the maximal independent set produced by the random greedy algorithm is not the only quantity of interest. We would also like to understand some of the structural properties of this set; in particular, what other properties of the binomial random set  $S(i)$  are shared by  $I(i)$ ?

For example, the lower bounds on  $R(s, t)$  mentioned above follow from the fact that the independence number of the graph produced by the  $K_s$ -free process is essentially the same as the independence number of the corresponding  $G_{n,p}$ . There has been extensive study of the number of copies of a fixed graph  $K$  that does not contain  $H$  as a subgraph in the graph produced by the  $H$ -free process [6, 14, 26, 31]. It turns out that the number of copies of such a graph  $K$  is roughly the same as in the corresponding  $G_{n,p}$ . Our next result is an extension of this fact to our general hypergraph setting.

Fix  $s$  and a  $s$ -uniform hypergraph  $\mathcal{G}$  on vertex set  $V$  (i.e. the same vertex set as the hypergraph  $\mathcal{H}$ ). We let  $X_{\mathcal{G}}(i)$  be the number of edges in  $\mathcal{G}$  that are contained in the independent set produced at the  $i^{\text{th}}$  step of the random greedy process on  $\mathcal{H}$ . Set  $p = p(i) = i/N$  and let  $i_{\max}$  be the lower bound (2) on the size of the independent set given by the random greedy algorithm given in Theorem 1.1.

**Theorem 1.2.** *If no edge of  $\mathcal{G}$  contains an edge of  $\mathcal{H}$ ,  $i < i_{\max}$  is fixed,  $|\mathcal{G}|p^s \rightarrow \infty$  and  $\Delta_a(\mathcal{G}) = o(p^a |\mathcal{G}|)$  for  $a = 1, \dots, s-1$  then*

$$X_{\mathcal{G}}(i) = |\mathcal{G}|p^s(1 + o(1)).$$

*with high probability.*

Note that Theorem 1.2 does not claim the conclusion holds for all  $i$  simultaneously. The proof is given in Section 5.

We believe that Theorems 1.1 and 1.2 may have multiple applications, most notably in the context of the  $H$ -free process where  $H$  is a  $k$ -uniform hypergraph (i.e.  $k \geq 3$  and our vertex set is  $V = \binom{[n]}{k}$ ). In Section 3 we discuss applications of Theorem 1.1 to hypergraph-free processes and note that in this context Theorem 1.1 gives new lower bounds on some hypergraph Turán problems. We also outline one other application: a lower bound on the number of steps in the  $k$ -AP-free process. This process forms a  $k$ -AP-free subset of  $\mathbb{Z}_N$  by adding elements chosen uniformly at random one at a time subject to the condition that no  $k$ -term arithmetic progression is formed. Details and discussion are given in the following Section.

## 2 The $k$ -AP-free process and Gowers uniformity norm

In this Section we address a question regarding  $k$ -term arithmetic progressions and the Gowers uniformity norm. The question was asked, independently, by Gowers [15] and Green [16] (for further discussion see Conlon, Fox and Zhao [10]). The Gowers  $U^d$  norm of  $f : \mathbb{Z}_N \rightarrow \mathbb{R}$  is

$$\|f\|_{U^d} := \left[ \frac{1}{N^{d+1}} \sum_{x \in \mathbb{Z}_N, h \in \mathbb{Z}_N^d} \prod_{\omega \in \{0,1\}^d} f(x + h \cdot \omega) \right]^{1/2^d}. \quad (5)$$

Given  $A \subset \mathbb{Z}_N$  we define a real-valued function  $\nu_A = \frac{N}{|A|} 1_A$ . Motivated by the study of the relationship between the Gowers norm and the distribution of arithmetic progressions in subsets of  $\mathbb{Z}_N$  (see Section 4 in [15]), Gowers and Green ask if there exists a function  $s(k)$  such that  $\|\nu_A - 1\|_{U^{s(k)}} = o(1)$  implies that  $A$  contains a  $k$ -term arithmetic progression.

Consider the  $k$ -AP-free process on  $\mathbb{Z}_N$  (the set of integers modulo  $N$ ), where  $N$  is prime. This process is an instance of the random greedy independent set algorithm on the  $k$ -uniform,  $k(N-1)$ -regular hypergraph  $\mathcal{H}_k$  which has vertex set  $\mathbb{Z}_N$  and edge set consisting of all  $k$ -term arithmetic progressions. We apply Theorems 1.1 and 1.2 to prove the following.

**Corollary 2.1.** *Let  $k, d$  be fixed integers such that  $2^{d-1} = k-1$ . Let  $N$  be prime. With high probability the  $k$ -AP-free process produces a set  $I(i_{\max}) \subseteq \mathbb{Z}_N$  such that*

$$\|\nu_I - 1\|_{U^d} = o(1).$$

Of course, the set  $I$  contains no  $k$ -term arithmetic progression. So we conclude that if the function  $s(k)$  exists then it satisfies  $s(k) > 1 + \log_2(k-1)$ . The remainder of this Section is a proof of Corollary 2.1.

We begin by noting that  $\mathcal{H}_k$  satisfies the conditions required for an application of Theorem 1.1. The condition on  $\Delta_a$  follows from the fact that any 2 elements of  $\mathbb{Z}_N$  are in at most  $k^2$  edges. Furthermore, for any  $v, v' \in \mathbb{Z}_N$  there are at most  $k^6$  pairs of edges  $e, e'$  such that  $v \in e, v' \in e'$ , and  $|e \cap e'| \geq 2$ . (Observe that setting the positions of 2 vertices in  $e \cap e'$  and  $v, v'$  in the two arithmetic progressions  $e$  and  $e'$  introduces a pair of linear equations that the differences for  $e$  and  $e'$  satisfy. This determines these differences uniquely.) Thus, we also have the desired condition on  $\Gamma$ . We conclude that with high probability the  $k$ -AP-free process produces a  $k$ -AP free set  $I$  of size

$$\Omega\left(N^{\frac{k-2}{k-1}} \log^{\frac{1}{k-1}} N\right).$$

In order to compute  $\|\nu_I - 1\|_{U^d}$  we consider the hypergraph  $\mathcal{G}$  of ‘ $d$ -cubes.’ For  $x \in \mathbb{Z}_N$  and  $h \in \mathbb{Z}_N^d$  define

$$e_{x,h} = \left\{ x + \omega \cdot h : \omega \in \{0,1\}^d \right\}.$$

We say that  $h \in \mathbb{Z}_N^d$  has *no coincidences* if the elements of  $\{\omega \cdot h : \omega \in \{-1, 0, 1\}^d\}$  are distinct. In other words,  $h$  has no coincidences if

$$\left| \left\{ \omega \cdot h : \omega \in \{-1, 0, 1\}^d \right\} \right| = 3^d.$$

We then define  $\mathcal{G}$  to be the hypergraph of  $d$ -cubes that have no coincidences; that is, we set

$$\mathcal{G} = \{e_{x,h} : h \text{ has no coincidences}\}.$$

We now establish the conditions required for an application of Theorem 1.2 to the number of edges of  $\mathcal{G}$  that appear in  $I$ . Note that  $|\mathcal{G}| = (1 + o(1))N^{d+1}$ . As  $p = |I|/N = \tilde{\Theta}(N^{-\frac{1}{k-1}})$ , we have

$$p^{2d} |\mathcal{G}| = \tilde{\Theta}(N^{-2} N^{d+1}) \rightarrow \infty.$$

Since we do not allow  $d$ -cubes with coincidences, no element of  $\mathcal{G}$  contains a  $k$ -AP (If  $h \cdot y - h \cdot x = h \cdot z - h \cdot y$  then  $h \cdot (y - x)$  coincides with  $h \cdot (z - y)$ .) The degree condition needed for an application of Theorem 1.2 is an easy consequence of the following Lemma.

**Lemma 2.2.** *For  $1 \leq a \leq 2^d$ , any set of  $a$  vertices is contained in at most  $O(N^{d - \lceil \log_2 a \rceil})$  edges of  $\mathcal{G}$ .*

*Proof.* First, note that there are  $(2^d)_a$  ways to specify which element of  $\{0, 1\}^d$  corresponds to each vertex in our set of  $a$  given vertices. So, it suffices to prove the following: For any  $y_1, \dots, y_a \in \mathbb{Z}_N$ , not necessarily distinct, and  $\omega_1, \dots, \omega_a \in \{0, 1\}^d$ , distinct, there are at most  $O(N^{d - \lceil \log_2 a \rceil})$  choices of  $x \in \mathbb{Z}_N$  and  $h \in \mathbb{Z}_N^d$  such that  $y_j = x + \omega_j \cdot h$  for  $j = 1, \dots, a$ .

We prove this statement by induction on  $d$ . Note that this statement is trivial if  $a = 1$ . This dispatches the base case ( $d = 1$ ) and allows us to hence forth consider  $a \geq 2$ .

WLOG assume that the Hamming distance between  $\omega_1$  and  $\omega_2$  is minimal among distances between pairs of vectors from  $\omega_1, \dots, \omega_a$ , and suppose that  $L \subset \{1, \dots, d\}$  is the set of coordinates at which  $\omega_1$  and  $\omega_2$  differ. Then there are  $O(N^{|L|-1})$  ways to specify the coordinates  $h_\ell$  for  $\ell \in L$  which are consistent with  $y_1 = x + \omega_1 \cdot h$  and  $y_2 = x + \omega_2 \cdot h$ .

Now, consider one of the ways to set the coordinates  $L$ . We may view the remainder of the embedding as a lower dimensional cube, with  $d' = d - |L|$  and  $a' \geq \frac{1}{2}a$  (since if there were three vectors  $\omega_j, \omega_{j'}, \omega_{j''}$  that only differed in coordinates of  $L$ , then two of them would have Hamming distance less than  $|L|$ , contradicting the fact that the Hamming distance from  $\omega_1$  to  $\omega_2$  is minimal).

Appealing to the induction hypothesis, altogether there are

$$O\left(N^{|L|-1} \cdot N^{(d-|L|) - \lceil \log_2(\frac{1}{2}a) \rceil}\right) = O\left(N^{d - \lceil \log_2 a \rceil}\right)$$

possible  $x, h$ .

□

We are now prepared to compute the uniformity norm for  $I = I(i_{\max})$ . Recall that  $p = \frac{|I|}{N}$ . Note that  $\nu_I = \frac{1}{p}1_I$ . We calculate  $\|\nu_I - 1\|_{U^d}$  by first considering the  $h \in \mathbb{Z}_N^d$  that have no coincidences. We will see below that the contribution from  $h$  with the property that  $\{h \cdot \omega : \omega \in \{-1, 0, 1\}^d\}$  are not all distinct is negligible.

Consider the number of edges of the hypergraph  $\mathcal{G}$  that are contained in  $I$ . As we note above, we can apply Theorem 1.2 to get an estimate for this number. Now note that if the hypotheses of Theorem 1.2 hold for  $\mathcal{G}$  then they also hold for the shadow of  $\mathcal{G}$  on the  $x^{\text{th}}$  level (i.e. the hypergraph on vertex set  $V$  whose edges are all sets of size  $x$  contained in an edge of  $\mathcal{G}$ ). We conclude that for each  $0 \leq x \leq 2^d$ , w.h.p. the number of edges of  $\mathcal{G}$  with exactly  $x$  vertices in  $I$  is

$$(1 + o(1))N^{d+1} \binom{2^d}{x} p^x.$$

Thus to compute  $\|\nu_I - 1\|_{U^d} = \|\frac{1}{p}1_I - 1\|_{U^d}$  using (5), we start with the sum of the terms corresponding to edges of  $\mathcal{G}$ . We get that

$$\begin{aligned} \sum_{e \in \mathcal{G}} \left(\frac{1}{p} - 1\right)^{|e \cap I|} (-1)^{2^d - |e \cap I|} &= N^{d+1} \sum_{0 \leq x \leq 2^d} (1 + o(1)) \binom{2^d}{x} p^x \left(\frac{1}{p} - 1\right)^x (-1)^{2^d - x} \\ &= N^{d+1} [(1 - p) - 1]^{2^d} + o(N^{d+1}) \\ &= o(N^{d+1}) \end{aligned}$$

with high probability.

It remains to address the terms in (5) corresponding to  $h$  such that the values  $\{h \cdot \omega : \omega \in \{-1, 0, 1\}^d\}$  are not all distinct. Each such vector  $h$  defines a partition  $\mathcal{P}$  of  $\{-1, 0, 1\}^d$ : each part  $P \in \mathcal{P}$  is a maximal subset of  $\{-1, 0, 1\}^d$  with the property that the values  $h \cdot \omega$  are the same for all  $\omega \in P$ .

We compute the remaining contribution to (5) by summing over all possible partitions. All vectors  $h$  that define a given partition  $\mathcal{P}$  satisfy a system of linear equations: namely, we have  $h \cdot \omega = h \cdot \omega'$  for every pair  $\omega, \omega'$  in the same part of  $\mathcal{P}$ . Suppose these equations give the matrix equation  $Ah = 0$ , and assume  $A$  has rank  $a$ . Then there are  $O(N^{d-a+1})$  pairs  $x, h$  that respect this partition  $\mathcal{P}$ .

The partition  $\mathcal{P}$  of  $\{-1, 0, 1\}^d$  defines a partition  $\overline{\mathcal{P}}$  of  $\{0, 1\}^d$  by restriction. (Of course, we only interested in  $\omega \in \{0, 1\}^d$  when computing the contribution to the Gowers norm). We claim that each part  $P$  of  $\overline{\mathcal{P}}$  we have  $|P| \leq 2^a$ . To see this, let  $\omega_0 \in P$  where  $P \in \overline{\mathcal{P}}$  and note that for every  $\omega \in P$ , the row space of  $A$  contains  $\omega - \omega_0$ . But a subspace of  $\mathbb{Z}_N^d$  of dimension  $a$  can only intersect  $\{0, -1\}^y \times \{0, 1\}^{d-y}$  in at most  $2^a$  points. So the row space of  $A$  can only contain  $2^a$  vectors that are 0 or  $-1$  on the support of  $\omega_0$  and 0 or 1 otherwise. Thus,  $|P| \leq 2^a$ .

For the partition  $\mathcal{P}$ , there are  $O(N^{d-a+1})$  pairs  $x, h$  that agree with  $\mathcal{P}$ . Fix such a pair  $x, h$ , and a collection of parts  $\mathcal{S} \subset \overline{\mathcal{P}}$ . Consider the event  $\mathcal{E}_{\mathcal{S}}$  that the images (under the map  $\varphi : \{0, 1\}^d \rightarrow \mathbb{Z}_N$  defined by  $\varphi(\omega) = x + h \cdot \omega$ ) of the parts of  $\mathcal{S}$  are in  $I$ , but none of the image of parts of  $\overline{\mathcal{P}} \setminus \mathcal{S}$  are in  $I$ . By a simple first moment calculation (using Lemma 5.1), we have

$$\mathbb{P}[\mathcal{E}_{\mathcal{S}}] = O(p^{|\mathcal{S}|}).$$

The sum of the terms in  $\|\nu_I - 1\|_{U^d}$  corresponding to pairs  $x, h$  such that  $h$  respects the partition  $\mathcal{P}$  is

$$\begin{aligned} O\left(N^{d-a+1} \sum_{\mathcal{S} \subset \overline{\mathcal{P}}} p^{|\mathcal{S}|} \left(\frac{1}{p} - 1\right)^{|\cup_{P \in \mathcal{S}} P|}\right) \\ = O\left(N^{d-a+1} p^{|\overline{\mathcal{P}}| - 2^d}\right) = O\left(N^{d-a+1} \cdot p^{2^{d-a} - 2^d}\right) = o\left(N^{d+1}\right), \end{aligned}$$

where we use  $k-1 = 2^{d-1}$  in the last equation. As there are only finitely many partitions  $\mathcal{P}$ , the proof of Corollary 2.1 is complete.

### 3 Hypergraph Turán

Consider the  $H$ -free process, where  $H$  is a  $k$ -uniform hypergraph ( $k \geq 2$ ) with vertex set  $V_H$  and edge set  $E_H$ . Set  $v_H = |V_H|$  and  $e_H = |E_H|$ . Recall that the  $H$ -free process is the same as random greedy independent set on the hypergraph  $\mathcal{H}_H$  that has vertex set  $\binom{[n]}{k}$  and edge set consisting of all copies of  $H$ . We say that  $H$  is strictly  $k$ -balanced if and only if

$$\frac{e_H[W] - 1}{|W| - k} < \frac{e_H - 1}{v_H - k} \quad \text{for all } W \subsetneq V_H \text{ such that } |W| > k, \quad (6)$$

where  $H[W]$  is the subgraph of  $H$  induced by  $W$ . We claim that  $\mathcal{H}_H$  satisfies (1) if and only if  $H$  is strictly  $k$ -balanced. To see this, note that in this setting we have  $r = e_H$ ,  $N = \Theta(n^k)$ ,  $D = \Theta(n^{v_H-k})$ . For  $a \geq 2$  let  $v_a$  be the minimum number of vertices of  $H$  spanned by a set of  $a$  edges of  $H$ . We have

$$\Delta_a(\mathcal{H}_H) = \Theta(n^{v_H-v_a}) = \Theta\left(n^{(v_H-k)\left[1-\frac{v_a-k}{v_H-k}\right]}\right)$$

and

$$D^{\frac{e_H-a}{e_H-1}} = \Theta\left(n^{\frac{(v_H-k)(e_H-a)}{e_H-1}}\right) = \Theta\left(n^{(v_H-k)\left[1-\frac{a-1}{e_H-1}\right]}\right).$$

Thus, we see that  $\mathcal{H}_H$  satisfies (1) if and only if  $(v_a - k)/(v_H - k) > (a - 1)/(e_H - 1)$  for all  $a \geq 2$ , which holds if and only if  $H$  is strictly  $k$ -balanced.

Therefore, Theorem 1.1 implies lower bounds on the Turán numbers of  $k$ -uniform, strictly  $k$ -balanced hypergraphs.

**Corollary 3.1.** *If  $H$  is a  $k$ -uniform, strictly  $k$ -balanced hypergraph with  $v_H$  vertices and  $e_H$  edges, then*

$$ex(n, H) = \Omega \left( n^{k - \frac{v_H - k}{e_H - 1}} \log^{\frac{1}{e_H - 1}} n \right).$$

For general  $k$ -partite, strictly  $k$ -balanced hypergraphs, the bounds we give are best known. (Although for a few fixed hypergraphs better bounds are known; for example, see [17],[20].) Note that the complete  $k$ -partite  $k$ -uniform hypergraph  $K_{s_1, \dots, s_k}$  is strictly  $k$ -balanced so long as all  $s_i \geq 1$  and some  $s_i \geq 2$ . To see this, suppose  $W$  is a set of vertices that has  $w_i$  from the  $i^{th}$  part. Then condition (6) becomes

$$\frac{\prod w_i - 1}{\sum w_i - k} < \frac{\prod s_i - 1}{\sum s_i - k}$$

for all sequence of integers  $w_1, w_2, \dots, w_k$  such that  $1 \leq w_i \leq s_i$  for all  $i$ ,  $w_i > 1$  for some  $i$ , and  $w_i < s_i$  for some  $i$ . Straightforward multivariate calculus shows that the above inequality holds for all real numbers  $w_1, \dots, w_k$  such that  $1 \leq w_i \leq s_i$ , except for the case when  $w_i = 1$  for all  $i$ , and the case when  $w_i = s_i$  for all  $i$ .

## 4 Lower bound: Proof of Theorem 1.1

We use dynamic concentration inequalities to prove that carefully selected statistics remain very close to their expected trajectories throughout the process with high probability. Our main goal is to prove dynamic concentration of  $|V(i)|$ , which is the number of vertices that remain in the hypergraph. In order to achieve this goal, we also track the following variables: For every vertex  $v \in V(i)$  and  $\ell = 2, \dots, r$  define  $d_\ell(i, v) = d_\ell(v)$  to be the number of edges of cardinality  $\ell$  in  $\mathcal{H}(i)$  that contain  $v$ .

We employ the following conventions throughout this section. If we arrive at a hypergraph  $\mathcal{H}(i)$  that has edges  $e, e'$  such that  $e \subseteq e'$  then we remove  $e'$  from  $\mathcal{H}$ . Note that this has no impact in the process as the presence of  $e$  ensures that we never have  $e' \subset I(j)$ . For any variable  $X$  we use the notation  $\Delta X := X(i+1) - X(i)$  for the one step change in  $X$ . Since *every* expectation taken in this section is conditional on the first  $i$  steps of the algorithm, we suppress the conditioning. That is, we simply write  $\mathbb{E}[\cdot]$  instead of  $\mathbb{E}[\cdot \mid \mathcal{F}_i]$  where  $\mathcal{F}_0, \mathcal{F}_1, \dots$  is the natural filtration generated by the algorithm. If  $f$  and  $g$  are functions of  $N$  with the property that  $f$  is bounded by  $g$  times some poly-logarithmic function (in  $N$ ) then we write  $f = \tilde{O}(g)$ .

We begin by discussing the expected trajectories of the variables we track. Here we use the binomial random set  $S(i)$  as a guide. Recall that each vertex is in  $S(i)$ , independently, with probability  $p = p(i) = i/N$ . Let  $v$  be a fixed vertex. The expected number of edges  $e \in \mathcal{H}$  such that  $v \in e$  and  $e \setminus \{v\} \subseteq S(i)$  is

$$Dp^{r-1} = D \left( \frac{i}{N} \right)^{r-1} = t^{r-1}$$

where we parametrize time by setting  $t := \frac{D^{\frac{1}{r-1}}}{N} \cdot i$ . Thus, we set

$$q = q(t) := e^{-t^{r-1}}$$

and think of  $q$  as the probability that a vertex is in  $V(i)$ . So, we should have  $|V(i)| \approx q(t)N$  and  $d_\ell(v)$  should follow the trajectory

$$s_\ell(t) := D \binom{r-1}{\ell-1} q^{\ell-1} p^{r-\ell} = \binom{r-1}{\ell-1} D^{\frac{\ell-1}{r-1}} t^{r-\ell} q^{\ell-1}.$$

For the purpose of our analysis, we separate the positive contributions to  $d_\ell(v)$  from the negative contributions. We write  $d_r(v) = D - d_r^-(v)$ , and for  $\ell < r$  we write  $d_\ell(v) = d_\ell^+(v) - d_\ell^-(v)$ , where  $d_\ell^+(v), d_\ell^-(v)$  are non-negative variables which count the number of edges of cardinality  $\ell$  containing  $v$  that are created and destroyed, respectively, through the first  $i$  steps of the process. We define

$$s_\ell^+(t) := D^{-\frac{1}{r-1}} \int_0^t \frac{\ell s_{\ell+1}(\tau)}{q(\tau)} d\tau \quad s_\ell^-(t) := D^{-\frac{1}{r-1}} \int_0^t \frac{(\ell-1)s_\ell(\tau)s_2(\tau)}{q(\tau)} d\tau,$$

and claim that should have  $d_\ell^\pm \approx s_\ell^\pm$ . Note that  $s_\ell$  satisfies the differential equation

$$s'_\ell = \frac{\ell s_{\ell+1} - (\ell-1)s_\ell s_2}{q} = (s_\ell^+)' - (s_\ell^-)'$$

and so  $s_\ell = s_\ell^+ - s_\ell^-$ . The choice of  $s_\ell^\pm$  is natural in light of the usual mechanism for establishing dynamic concentration and the observation that we have

$$\mathbb{E}[\Delta d_\ell^+] \approx \frac{1}{|V(i)|} \cdot \ell d_{\ell+1} \quad \mathbb{E}[\Delta d_\ell^-] \approx \frac{1}{|V(i)|} \cdot (\ell-1) d_\ell d_2.$$

In addition to our dynamic concentration estimates, we need some auxiliary information about the evolving hypergraph  $\mathcal{H}(i)$ .

**Definition 1** (Degrees of Sets). *For a set of vertices  $A$  of at least 2 vertices, let  $d_{A \uparrow b}(i)$  be the number of edges of size  $b$  containing  $A$  in  $\mathcal{H}(i)$ .*

**Definition 2** (Codegrees). *For a pair of vertices  $v, v'$ , let  $c_{a,a' \rightarrow k}(v, v', i)$  be the number of pairs of edges  $e, e'$ , such that  $v \in e \setminus e', v' \in e' \setminus e$ ,  $|e| = a, |e'| = a'$  and  $|e \cap e'| = k$  and  $e, e' \in \mathcal{H}(i)$ .*

We do not establish dynamic concentration for these variables, but we only need relatively crude upper bounds.

In order to state our results precisely, we introduce a stopping time. Set

$$i_{\max} := \zeta N D^{-\frac{1}{r-1}} \log^{\frac{1}{r-1}} N \quad \text{and} \quad t_{\max} := \frac{D^{\frac{1}{r-1}}}{N} \cdot i_{\max} = \zeta \log^{\frac{1}{r-1}} N,$$

where  $\zeta > 0$  is a constant (the choice of this constant is discussed below). Note that we have

$$q(t_{\max}) = N^{-\zeta^{r-1}}. \quad (7)$$

Define the stopping time  $T$  as the minimum of  $i_{\max}$  and the first step  $i$  such that any of the following conditions fails to hold:

$$|V(i)| \in Nq \pm ND^{-\delta} f_v \quad (8)$$

$$d_\ell^\pm(v) \in s_\ell^\pm \pm D^{\frac{\ell-1}{r-1}-\delta} f_\ell \quad \text{for } \ell = 2, \dots, r \text{ and all } v \in V(i) \quad (9)$$

$$d_{A \uparrow b} \leq D_{a \uparrow b} \quad \text{for } 2 \leq a < b \leq r \text{ and all } A \in \binom{V(i)}{a} \quad (10)$$

$$c_{a,a' \rightarrow k}(v, v') \leq C_{a,a' \rightarrow k} \quad \text{for all } v, v' \in V(i) \quad (11)$$

where  $\delta > 0$  is a constant and  $f_v, f_2, \dots, f_r$  are functions of  $t$  (which will stay small enough so that the error terms are always little-o of the main terms) and  $D_{a \uparrow b}$  and  $C_{a,a' \rightarrow k}$  are functions of  $D$  (but not  $t$ ) that satisfy

$$D_{a \uparrow b} \leq D^{\frac{b-a}{r-1}-\frac{\epsilon}{2}}$$

$$C_{a,a' \rightarrow k} \leq D^{\frac{a+a'-k-2}{r-1}-\frac{\epsilon}{2}}.$$

All of these parameters are specified explicitly below.

We prove Theorem 1.1 by showing that  $\mathbb{P}(T < i_{\max}) < \exp\{-N^{\Omega(1)}\}$ . We break the proof into two parts. We first establish the crude bounds, namely (10) and (11), in Section 4.1. We then turn to the dynamic concentration inequalities (8) and (9) in Section 4.2.

The constants  $\zeta, \delta$  are chosen so that

$$\zeta \ll \delta \ll \epsilon,$$

in the sense that  $\delta$  is chosen to be sufficiently small with respect to  $\epsilon$ , and  $\zeta$  is chosen to be sufficiently small with respect to  $\delta$ . The martingales that we consider below are stopped in the sense that when we define a sequence  $Z(i)$  we in fact work with  $Z(i \wedge T)$ . Thus we can assume that the bounds (8)-(11) always hold. The martingales that depend on a fixed vertex  $v$  (or a fixed sets of vertices  $A$ ) are also *frozen* in the sense that we set  $Z(i) = Z(i-1)$  if the vertex  $v$  (or some vertex in the fixed set  $A$ ) is not in  $V(i)$ .

## 4.1 Crude bounds

Define

$$D_{a \uparrow b} := D^{\frac{b-a}{r-1}-\epsilon+2(r-b)\lambda}$$

$$C_{a,a' \rightarrow k} := 2^r D^{\frac{a+a'-k-2}{r-1}-\epsilon+(2r-2k-2)\lambda}$$

where  $\lambda = \epsilon/4r$ . Throughout this section we use the bound  $|V(i)| > ND^{-\lambda}$ , which follows from (7) and the fact that we may set  $\zeta > 0$  sufficiently small relative to  $\epsilon$ .

**Lemma 4.1.** *Let  $2 \leq a < b \leq r$ .*

$$\mathbb{P} \left( \exists i \leq T \text{ and } A \in \binom{V(i)}{a} \text{ such that } d_{A \uparrow b}(i) \geq D_{a \uparrow b} \right) \leq \exp \left\{ -N^{\Omega(1)} \right\}.$$

*Proof.* We go by reverse induction on  $b$ . Note that if  $b = r$  then the desired bound follows immediately from the condition on  $\Delta_a(\mathcal{H})$  assumed in the statement of Theorem 1.1.

Let  $b < r$  and consider a fixed  $A \in \binom{V}{a}$ . For  $0 \leq j \leq D_{a+1 \uparrow b+1}$ , let  $N_j(i)$  be the number of vertices in  $V(i)$  (but not in  $A$ ) that appear in  $j$  of the edges counted by  $d_{A \uparrow b+1}(i)$ . Note that clearly  $\sum N_j(i) = |V(i)|$ , and  $\sum j N_j(i) = (b+1-a)d_{A \uparrow b+1} \leq (b+1-a)D_{a \uparrow b+1}$ .

Then (w.r.t. the filtration  $\mathcal{F}_i$ )  $\Delta d_{A \uparrow b}(i)$  is stochastically dominated by  $\Delta X(i)$ , where  $X(i)$  is a variable such that  $X(0) = 0$  and

$$\mathbb{P}(\Delta X = j) = \frac{N_j(i)}{|V(i)|}$$

We will bound  $d_{A \uparrow b}(i)$  by appealing to this stepwise stochastic domination, and bound  $X(i)$  using the following lemma due to Freedman [13]:

**Lemma 4.2** (Freedman). *Let  $Y(i)$  be a supermartingale, with  $\Delta Y(i) \leq C$  for all  $i$ , and  $V(i) := \sum_{k \leq i} \text{Var}[\Delta Y(k) | \mathcal{F}_k]$ . Then*

$$\mathbb{P}[\exists i : V(i) \leq v, Y(i) - Y(0) \geq d] \leq \exp \left( -\frac{d^2}{2(v + Cd)} \right).$$

To apply the lemma, we calculate

$$\mathbb{E}[\Delta X] = \frac{1}{|V(i)|} \sum j N_j \leq \frac{1}{ND^{-\lambda}} (b+1-a) D_{a \uparrow b+1} \leq \frac{r}{N} D^{\frac{b-a+1}{r-1} - \epsilon + (2r-2b-1)\lambda}.$$

Thus if we define

$$Y(i) := X(i) - \frac{r}{N} D^{\frac{b-a+1}{r-1} - \epsilon + (2r-2b-1)\lambda} \cdot i$$

then  $Y(i)$  is a supermartingale. Now

$$\begin{aligned} \text{Var}[\Delta Y] &= \text{Var}[\Delta X] \leq \mathbb{E}[(\Delta X)^2] = \frac{1}{|V(i)|} \sum j^2 N_j(i) \\ &\leq \frac{D_{a+1 \uparrow b+1}}{|V(i)|} \sum j N_j \leq \frac{D_{a+1 \uparrow b+1}}{|V(i)|} \cdot r D_{a \uparrow b+1} \leq \frac{r}{N} D^{\frac{2b-2a+1}{r-1} - 2\epsilon + (4r-4b-3)\lambda} \end{aligned}$$

So we apply Lemma 4.2 with

$$v = (\log N) D^{\frac{2b-2a}{r-1} - 2\epsilon + (4r-4b-3)\lambda},$$

recalling (4), and  $C = D_{a+1 \uparrow b+1} = D^{\frac{b-a}{r-1} - \epsilon + (2r-2b-2)\lambda}$  to conclude that we have

$$P \left[ Y(i) \geq D^{\frac{b-a}{r-1} - \epsilon + (2r-2b-1)\lambda} \right] \leq \exp \left\{ -N^{\Omega(1)} \right\}.$$

This suffices to complete the proof (applying the union bound over all choices of the set  $A$ ).  $\square$

**Lemma 4.3.** *Let  $2 \leq a, a' \leq r$  and  $1 \leq k < a, a'$  be fixed.*

$$\mathbb{P}(\exists i \leq T \text{ and } v, v' \in V(i) \text{ such that } c_{a,a' \rightarrow k}(v, v', i) \geq C_{a,a' \rightarrow k}) \leq \exp \left\{ -N^{\Omega(1)} \right\}.$$

*Proof.* We first note that Lemma 4.1 implies Lemma 4.3 except in the case  $a = a' = k+1$ . To see this, suppose  $a' > k+1$  and let  $v, v'$  be any two vertices. Then we have

$$c_{a,a' \rightarrow k}(v, v') \leq d_a(v) \cdot \binom{a}{k} \cdot D_{k+1 \uparrow a'} \leq D^{\frac{a-1}{r-1}} \cdot 2^r \cdot D^{\frac{a'-(k+1)}{r-1} - \epsilon + 2(r-a')\lambda},$$

which gives the desired bound.

So we restrict our attention to the case  $a = a' = k+1$ . We again proceed by induction, with the base case following immediately from the condition on  $\Gamma(\mathcal{H})$ . Note that  $c_{k+1,k+1 \rightarrow k}(v, v', i)$  can increase in size only when the algorithm chooses a vertex contained in the intersection of a pair of edges from  $c_{k+2,k+2 \rightarrow k+1}(v, v', i)$ , or when the algorithm chooses the vertex not contained in the intersection of a pair of edges from  $c_{k+2,k+1 \rightarrow k}(v, v', i)$  or  $c_{k+1,k+2 \rightarrow k}(v, v', i)$ . Also, on steps when  $c_{k+1,k+1 \rightarrow k}(v, v', i)$  does increase, it increases by at most  $2D_{2 \uparrow k+2} + D_{2 \uparrow k+1} \leq 3D^{\frac{k}{r-1} - \epsilon + (2r-2k-4)\lambda}$ .

For  $0 \leq j \leq 3D^{\frac{k}{r-1} - \epsilon + (2r-2k-4)\lambda}$ , let  $N_j(i)$  be the number of vertices that, if chosen, would increase  $c_{k+1,k+1 \rightarrow k}(v, v')$  by  $j$ . Note that  $\sum N_j(i) = |V(i)|$  while

$$\sum j N_j(i) \leq C_{k+2,k+2 \rightarrow k+1} + C_{k+2,k+1 \rightarrow k} + C_{k+1,k+2 \rightarrow k} \leq 3 \cdot 2^r D^{\frac{k+1}{r-1} - \epsilon + (2r-2k-2)\lambda}.$$

Then (w.r.t. the filtration  $\mathcal{F}_i$ )  $\Delta c_{k+1,k+1 \rightarrow k}(v, v')(i)$  is stochastically dominated by  $\Delta X(i)$ , where  $X(i)$  is a variable such that  $X(0) = 0$  and  $Pr(\Delta X = j) = \frac{N_j(i)}{|V(i)|}$ .

We apply Lemma 4.2 to bound  $X$ . We define

$$Y(i) := X(i) - \frac{3 \cdot 2^r}{N} D^{\frac{k+1}{r-1} - \epsilon + (2r-2k-1)\lambda} \cdot i$$

and note that  $Y(i)$  is a supermartingale. Now

$$\begin{aligned} Var[\Delta Y] &= Var[\Delta X] \leq \mathbb{E}[(\Delta X)^2] = \frac{1}{|V(i)|} \sum j^2 N_j(i) \\ &\leq \frac{3D^{\frac{k}{r-1} - \epsilon + (2r-2k-4)\lambda}}{ND^{-\lambda}} \sum j N_j \leq \frac{9 \cdot 2^r}{N} D^{\frac{2k+1}{r-1} - 2\epsilon + (4r-4k-5)\lambda} \end{aligned}$$

Applying Lemma 4.2 with

$$v = (\log N) D^{\frac{2k}{r-1} - 2\epsilon + (4r-4k-5)\lambda},$$

recalling (4), and  $C = 3D^{\frac{k}{r-1} - \epsilon + (2r-2k-4)\lambda}$  we have

$$P \left[ Y(i) \geq D^{\frac{k}{r-1} - \epsilon + (2r-2k-2.1)\lambda} \right] \leq \exp \left\{ -N^{\Omega(1)} \right\}.$$

□

## 4.2 Dynamic concentration

Consider the sequences

$$\begin{aligned} Z_V &:= |V(i)| - Nq - ND^{-\delta} f_v \\ Z_\ell^+(v) &:= d_\ell^+(v) - s_\ell^+ - D^{\frac{\ell-1}{r-1} - \delta} f_\ell & \text{for } 2 \leq \ell \leq r-1 \\ Z_\ell^-(v) &:= d_\ell^-(v) - s_\ell^- - D^{\frac{\ell-1}{r-1} - \delta} f_\ell & \text{for } 2 \leq \ell \leq r \end{aligned}$$

We establish the upper bound on  $V(i)$  in (8) by showing that  $Z_V < 0$  for all  $i \leq T$  with high probability. Similarly, we establish the upper bounds on  $d_\ell^\pm(v)$  in (9) by showing that  $Z_\ell^\pm(v) < 0$  for all  $i \leq T$  with high probability. The lower bounds follow from the consideration of analogous random variables.

We begin by showing that the sequences  $Z_V$  and  $Z_\ell^\pm$  are supermartingales. We will see that each of these calculations imposes a condition (inequality) on the collection of error functions  $\{f_v\} \cup \{f_\ell \mid \ell = 2, \dots, r\}$  and their derivatives. These differential equations are the *variation equations*. We choose error functions that satisfy the variation equations after completing the expected change calculations. The functions will be chosen so that all error functions evaluate to 1 at  $t = 0$  and are increasing in  $t$ . After we establish that the sequences are indeed supermartingales, we use the fact that they have initial values that are negative and relatively large in absolute value. We complete the proof by applying martingale deviation inequalities to show that it is very unlikely for these supermartingales to ever be positive.

We start the martingale calculations with the variable  $Z_V$ . We treat this first case in some detail in an effort to illuminate our methods for the reader who is not familiar with these techniques. Let  $S_t = D^{\frac{1}{r-1}}/N$  and recall that  $t = i/S_t$ . (The quantity  $S_t$  is sometimes called the *time scaling*.) We write

$$\begin{aligned} \Delta Z_V &= (|V(i+1)| - |V(i)|) - N(q(t+1/S_t) - q(t)) \\ &\quad - ND^{-\delta}(f_v(t+1/S_t) - f_v(t)), \end{aligned}$$

and make use of the estimates

$$\begin{aligned} q(t+1/S_t) - q(t) &= \frac{q'(t)}{S_t} + O\left(\frac{q''}{S_t^2}\right) \\ f_v(t+1/S_t) - f_v(t) &= \frac{f'_v(t)}{S_t} + O\left(\frac{f''_v}{S_t^2}\right) \end{aligned}$$

where  $q''$  and  $f_v''$  are understood to be bounds on the second derivative that hold uniformly in the interval of interest. We will see that the main terms of  $\Delta|V(i)|$  and  $N(q(t+1/S_t) - q(t))$  cancel exactly. The second order terms from  $\Delta|V(i)|$  will be then be balanced by the  $ND^{-\delta}(f_v(t+1/S_t) - f_v(t))$  term. We now proceed with the explicit details.

Note that  $q' = -s_2 D^{-\frac{1}{r-1}}$  and that  $q''$  is  $q$  times a polynomial in  $t$ , and recall  $N = \Omega\left(D^{\frac{1}{r-1}+\epsilon}\right)$  (see (4)). We have

$$\begin{aligned}\mathbb{E}[\Delta Z_V] &= -\frac{1}{|V(i)|} \left[ \sum_{v \in V(i)} (d_2(v) + 1) \right] + s_2 - D^{\frac{1}{r-1}-\delta} f'_v \\ &\quad + \tilde{O}\left(\frac{D^{\frac{2}{r-1}}}{N} q\right) + O\left(\frac{D^{\frac{2}{r-1}-\delta}}{N} f_v''\right) \\ &\leq D^{\frac{1}{r-1}-\delta} [2f_2 - f'_v] + O\left(D^{\frac{1}{r-1}-\delta-\epsilon} f_v'' + D^{\frac{1}{r-1}-\epsilon+o(1)}\right)\end{aligned}$$

whence we derive the first variation equation:

$$f'_v > 2f_2. \quad (12)$$

While a condition along the lines the of (12) suffices, we impose a stronger condition in order to simplify our calculations. We will choose our error functions so that we have

$$f'_v > 3f_2. \quad (13)$$

So long as (13) holds and  $f_v'' = o(D^\epsilon)$ , the sequence  $Z_V$  is a supermartingale (note that we apply the fact that  $f_2 > 1$  here). We address the condition on  $f_v''$  below.

Now we turn to  $Z_\ell^+$ . (The reader familiar with the original analysis of the  $H$ -free process [6] should note that there is no ‘creation fidelity’ term here as, thanks to the convention that removes any edge that contains another edge, selection of a vertex in an edge  $e$  cannot close another vertex in the same edge.) For  $2 \leq \ell \leq r-1$  we have

$$\begin{aligned}\mathbb{E}[\Delta Z_\ell^+(v)] &= \frac{\ell d_{\ell+1}(v)}{|V(i)|} - \frac{\ell s_{\ell+1}}{Nq} - \frac{D^{\frac{\ell}{r-1}-\delta}}{N} f'_\ell \\ &\quad + \tilde{O}\left(\frac{D^{\frac{\ell+1}{r-1}}}{N^2}\right) + O\left(\frac{D^{\frac{\ell+1}{r-1}-\delta}}{N^2} f_\ell''\right) \\ &\leq \frac{\ell \left(s_{\ell+1} + 2D^{\frac{\ell}{r-1}-\delta} f_{\ell+1}\right)}{Nq - ND^{-\delta} f_v} - \frac{\ell s_{\ell+1}}{Nq} - \frac{D^{\frac{\ell}{r-1}-\delta}}{N} f'_\ell \\ &\quad + \tilde{O}\left(\frac{D^{\frac{\ell}{r-1}-\epsilon}}{N}\right) + O\left(\frac{D^{\frac{\ell}{r-1}-\delta-\epsilon}}{N} f_\ell''\right) \\ &\leq \frac{D^{\frac{\ell}{r-1}-\delta}}{N} \cdot \left[ 2\ell q^{-1} f_{\ell+1} + \ell \binom{r-1}{\ell} t^{r-\ell-1} q^{\ell-2} f_v - f'_\ell \right] \\ &\quad + \tilde{O}\left(\frac{D^{\frac{\ell}{r-1}-\epsilon}}{N}\right) + O\left(\frac{D^{\frac{\ell}{r-1}-\delta-\epsilon}}{N} f_\ell''\right)\end{aligned}$$

(note on the second line we use  $s_{\ell+1} = s_{\ell+1}^+ - s_{\ell+1}^-$ ) whence we derive the following variation equations for  $2 \leq \ell \leq r-1$ :

$$f'_\ell > 5\ell q^{-1} f_{\ell+1} \quad (14)$$

$$f'_\ell > 2\ell \binom{r-1}{\ell} t^{r-\ell-1} q^{\ell-2} f_v \quad (15)$$

So long as (14), (15) hold,  $\delta < \epsilon$  and  $f''_\ell = o(D^\epsilon)$  the sequence  $Z_\ell^+(v)$  is a supermartingale.

Finally, we consider  $Z_\ell^-$  for  $2 \leq \ell \leq r$ . The main term in the expected change of  $d_\ell^-(v)$  comes from the selection of vertices  $y$  for which there exists a vertex  $x$  such that  $\{y, x\} \in \mathcal{H}(i)$  and there is an edge  $e$  counted by  $d_\ell(v)$  such that  $x \in e$ . However, here we must also account for the convention of removing redundant edges. For a fixed edge  $e$  counted by  $d_\ell(v)$ , the selection of any vertex in the following sets results in the removal of  $e$  from this count:

$$\{y \in V(i) : \exists A \subset e \text{ such that } A \neq \{v\} \text{ and } A \cup \{y\} \in \mathcal{H}(i)\}.$$

(Note that this is essentially restating the convention that we remove edges that contain other edges). Together, the sums

$$\sum_{x \in e \setminus \{v\}} d_2(x) + \sum_{A \subset e, |A| \geq 2} d_{A \uparrow |A|+1}$$

count each  $y$  with the property that the choice of  $y$  causes the removal of  $e$  from the count  $d_\ell(v)$  at least once and at most  $O(1)$  many times. The number of  $y$  that are counted more than once in the first sum is at most  $\binom{\ell-1}{2} C_{2,2 \rightarrow 1}$ . Therefore we get the following estimate:

$$\mathbb{E} [\Delta d_\ell^-(v)] = \frac{1}{|V|} \left\{ \sum_{e \in d_\ell(v)} \sum_{u \in e \setminus \{v\}} d_2(u) + O \left( d_\ell \cdot \left[ C_{2,2 \rightarrow 1} + \sum_{k=2}^{\ell-1} D_{k \uparrow k+1} \right] \right) \right\}$$

(We note in passing that this estimate also takes into account the selection of vertices that cause  $v$  itself to be removed from the vertex set. In other words, this estimate also

takes freezing into account.) We have

$$\begin{aligned}
\mathbb{E} [\Delta Z_\ell^-(v)] &= \frac{1}{|V(i)|} \left\{ \sum_{e \in d_\ell(v)} \sum_{u \in e \setminus \{v\}} d_2(u) + O \left( d_\ell \cdot \left[ C_{2,2 \rightarrow 1} + \sum_{k=2}^{\ell-1} D_{k \uparrow k+1} \right] \right) \right\} \\
&\quad - \frac{(\ell-1)s_\ell \cdot s_2}{Nq} - \frac{D^{\frac{\ell}{r-1}-\delta}}{N} f'_\ell + \tilde{O} \left( \frac{D^{\frac{\ell+1}{r-1}}}{N^2} \right) + O \left( \frac{D^{\frac{\ell+1}{r-1}-\delta}}{N^2} f''_\ell \right) \\
&\leq \frac{(\ell-1) \left( s_\ell + 2D^{\frac{\ell-1}{r-1}-\delta} f_\ell \right) \left( s_2 + 2D^{\frac{1}{r-1}-\delta} f_2 \right)}{Nq - ND^{-\delta} f_v} - \frac{(\ell-1)s_\ell \cdot s_2}{Nq} - \frac{D^{\frac{\ell}{r-1}-\delta}}{N} f'_\ell \\
&\quad + \tilde{O} \left( \frac{D^{\frac{\ell+1}{r-1}}}{N^2} \right) + O \left( \frac{D^{\frac{\ell}{r-1}-\frac{\epsilon}{2}}}{N} q^{\ell-2} + \frac{D^{\frac{\ell+1}{r-1}-\delta}}{N^2} f''_\ell \right) \\
&\leq \frac{D^{\frac{\ell}{r-1}-\delta}}{N} \cdot \left[ (2+o(1))(\ell-1) \binom{r-1}{\ell-1} t^{r-\ell} q^{\ell-2} f_2 + 2(\ell-1)(r-1) t^{r-2} f_\ell \right. \\
&\quad \left. + (\ell-1)(r-1) \binom{r-1}{\ell-1} t^{2r-\ell-2} q^{\ell-2} f_v - f'_\ell \right] \\
&\quad + \tilde{O} \left( \frac{D^{\frac{\ell}{r-1}-\epsilon}}{N} \right) + O \left( \frac{D^{\frac{\ell}{r-1}-\frac{\epsilon}{2}}}{N} q^{\ell-2} + \frac{D^{\frac{\ell}{r-1}-\delta-\epsilon}}{N} f''_\ell \right)
\end{aligned}$$

whence we derive the following variation equations for  $2 \leq \ell \leq r$ :

$$f'_\ell > 7(\ell-1) \binom{r-1}{\ell-1} t^{r-\ell} q^{\ell-2} f_2 \quad (16)$$

$$f'_\ell > 6(\ell-1)(r-1) t^{r-2} f_\ell \quad (17)$$

$$f'_\ell > 3(\ell-1)(r-1) \binom{r-1}{\ell-1} t^{2r-\ell-2} q^{\ell-2} f_v \quad (18)$$

So long as (16), (17), (18) hold and  $\epsilon/2 > \delta$  and  $f''_\ell = o(D^\epsilon)$  the sequence  $Z_\ell^-(v)$  is a supermartingale. (We will see below that  $q^\ell f_2 > 1$ .)

We satisfy the variation equations (13), (14), (15), (16), (17), (18) by setting the error functions to have the form

$$f_\ell = \left(1 + t^{r-\ell+2}\right) \cdot \exp(\alpha t + \beta t^{r-1}) \cdot q^\ell$$

$$f_v = (1 + t^2) \cdot \exp(\alpha t + \beta t^{r-1}) \cdot q^2$$

for some constants  $\alpha$  and  $\beta$  depending only on  $r$ . Note that (dropping some terms) we have

$$f'_\ell \geq \left[ \alpha + (\beta - \ell)(r-1) t^{2r-\ell} \right] \cdot \exp(\alpha t + \beta t^{r-1}) \cdot q^\ell$$

$$f'_v \geq [\alpha + (\beta - 2)(r-1) t^r] \cdot \exp(\alpha t + \beta t^{r-1}) \cdot q^2.$$

Note that for this choice of functions, all variation equations have the property that both sides of the equation have the same exponential term. It remains to compare the

polynomial terms; in each case it is clear that we get the desired inequality by choosing  $\alpha$  and  $\beta$  to be sufficiently large (as functions of  $r$ ). We get the desired conditions on second derivatives by choosing  $\zeta$  sufficiently small (recall that we are free to choose  $\zeta$  arbitrarily small).

We complete the proof by applying martingale variation inequalities to prove that  $Z_V$  and  $Z_\ell^\pm$  remain negative with high probability. We will apply the following lemmas (which both follow from Hoeffding [18]):

**Lemma 4.4.** *Let  $X_i$  be a supermartingale such that  $|\Delta X| \leq c_i$  for all  $i$ . Then*

$$\mathbb{P}(X_m - X_0 > d) \leq \exp \left( -\frac{d^2}{2 \sum_{i \leq m} c_i^2} \right)$$

**Lemma 4.5.** *Let  $X_i$  be a supermartingale such that  $-N \leq \Delta X \leq \eta$  for all  $i$ , for some  $\eta < \frac{N}{10}$ . Then for any  $d < \eta m$  we have*

$$\mathbb{P}(X_m - X_0 > d) \leq \exp \left( -\frac{d^2}{3m\eta N} \right)$$

For our upper bound on  $|V(i)|$  we apply Lemma 4.4 to the supermartingale  $Z_V(i)$ . Note that if the vertex  $v$  is inserted to the independent set at step  $i$  then  $\Delta V = -1 - d_2(v) \in s_2 \pm D^{\frac{1}{r-1}-\delta} f_2$ . As  $q' = -D^{-\frac{1}{r-1}} s_2$  we have

$$\begin{aligned} \Delta Z_V &= O \left( D^{\frac{1}{r-1}-\delta} (f_2 + f'_v) \right) + O \left( \frac{D^{\frac{2}{r-1}-\delta}}{N} f''_v \right) + \tilde{O} \left( \frac{D^{\frac{2}{r-1}}}{N} \right) \\ &= O \left( D^{\frac{1}{r-1}-\delta} (f_2 + f'_v) \right). \end{aligned}$$

As we have  $Z_V(0) = -ND^{-\delta}$ , the probability that  $Z_V$  is positive at step  $T$  is at most

$$\begin{aligned} \exp \left\{ -\tilde{\Omega} \left( \frac{(ND^{-\delta})^2}{ND^{-\frac{1}{r-1}} \cdot \left[ D^{\frac{1}{r-1}-\delta} (f_2 + f'_v) \right]^2} \right) \right\} &\leq \exp \left\{ -\tilde{\Omega} (D^\epsilon (f_2 + f'_v)^{-2}) \right\} \\ &\leq \exp \left\{ -N^{\Omega(1)} \right\}, \end{aligned}$$

so long as  $\zeta$  is sufficiently small. Note that we have used the notation  $\tilde{\Omega}(\cdot)$ : If  $f$  and  $g$  are functions of  $N$  such that  $f$  is bounded above by  $g$  times some poly-logarithmic factor we write  $f = \tilde{\Omega}(g)$ . Also note that  $D > N^\epsilon$  (see (4)) is used to get the last expression above.

For our bound on  $d_\ell^+(v)$  we apply Lemma 4.5 to the supermartingale  $Z_\ell^+(v)$ . Note that we have

$$\Delta Z_\ell^+(v) < D_{2\uparrow\ell+1} \leq D^{\frac{\ell-1}{r-1}-\frac{\epsilon}{2}}$$

and note that the upper bound is not affected by the convention of removing edges containing other edges, since that would only decrease the number of new edges. For a lower bound we have

$$\Delta Z_\ell^+(v) > -(s_\ell^+)' \cdot \frac{D^{\frac{1}{r-1}}}{N} + O\left((s_\ell^+)'' \cdot \frac{D^{\frac{2}{r-1}}}{N^2} + f_\ell' \cdot \frac{D^{\frac{1}{r-1}}}{N}\right) > -O\left(\frac{D^{\frac{\ell}{r-1}}}{N}\right).$$

We have  $Z_\ell^+(0) = -D^{\frac{\ell-1}{r-1}-\delta}$ , and the hypotheses of Lemma 4.5 hold since  $\frac{D^{\frac{\ell}{r-1}}}{N} = o(D^{\frac{\ell-1}{r-1}-\frac{\epsilon}{2}})$  and  $D^{\frac{\ell-1}{r-1}-\delta} = o(\frac{D^{\frac{\ell}{r-1}}}{N} \cdot i_{\max})$ . So the probability that  $Z_\ell^+(v)$  is positive at some step  $i \leq T$  is at most

$$\begin{aligned} \exp\left\{-\tilde{\Omega}\left(\frac{\left(D^{\frac{\ell-1}{r-1}-\delta}\right)^2}{ND^{-\frac{1}{r-1}} \cdot \frac{1}{N}D^{\frac{\ell}{r-1}} \cdot D^{\frac{\ell-1}{r-1}-\epsilon/2}}\right)\right\} \\ \leq \exp\left\{-\tilde{\Omega}\left(D^{\frac{\epsilon}{2}-2\delta}\right)\right\} \leq \exp\left\{-N^{\Omega(1)}\right\}. \end{aligned}$$

Note that we have use  $\delta < \epsilon/4$  to obtain the last expression.

For our bound on  $d_\ell^-(v)$  we apply Lemma 4.5 to the supermartingale  $Z_\ell^-$ . Note that

$$-O\left(\frac{D^{\frac{\ell}{r-1}}}{N}\right) < \Delta Z_\ell^-(v) < O\left(\sum_{1 \leq k \leq \ell-1} C_{\ell,k+1 \rightarrow k}\right) = O\left(D^{\frac{\ell-1}{r-1}-\frac{\epsilon}{2}}\right)$$

(and note here that the upper bound includes edges lost because they contain other edges). Thus, the rest of the calculation is the same as it was for  $d_\ell^+(v)$ .

## 5 Subgraph counts: Proof of Theorem 1.2

Here we apply the observation, due to Wolfowitz [31], that the classical second moment argument for subgraph counts can be applied in the context of the random greedy independent set process.

**Lemma 5.1.** *Fix a constant  $L$  and suppose  $\{v_1 \dots v_L\} \subset V$  does not contain an edge of  $\mathcal{H}$ . Then for all  $j \leq i_{\max}$  we have*

$$\mathbb{P}(\{v_1 \dots v_L\} \subset I(j)) = (j/N)^L \cdot (1 + o(1)).$$

*Proof.* Fix a permutation of this set of vertices, say  $u_1 \dots u_L$  after relabeling, and a list of steps of the algorithm  $i_1 < \dots < i_L \leq j$ . Let  $\mathcal{E}$  be the event that each  $u_k$  is chosen on step  $i_k$  for  $k = 1, \dots, L$ . Note that the event  $\mathcal{E}$  requires that vertex  $u_k$  remains in  $V(i)$  until step  $i_k - 1$ , and, in order to achieve this condition, the set  $\{v_1 \dots v_L\}$  can never contain an edge of  $\mathcal{H}(i)$ .

Let  $\mathcal{E}_i$  be the event that  $T > i$  and the first  $i$  steps of the algorithm are compatible with  $\mathcal{E}$ . Then we write

$$\mathbb{P}(\mathcal{E}_1) \prod_{i=2}^{i_L} \mathbb{P}(\mathcal{E}_i \mid \mathcal{E}_{i-1}) \leq \mathbb{P}(\mathcal{E}) \leq \mathbb{P}(\mathcal{E}_1) \prod_{j=2}^{i_L} \mathbb{P}(\mathcal{E}_j \mid \mathcal{E}_{j-1}) + \mathbb{P}(T \leq i_L).$$

If  $i = i_k$  then, conditioning on the first  $i - 1$  steps of the algorithm and the event  $\mathcal{E}_{i-1}$ , we have  $\mathbb{P}(\mathcal{E}_i) = \frac{1}{|V(i-1)|}$ , unless the selection of  $u_k$  triggers the stopping time  $T$ . Also recall that  $\mathbb{P}(T \leq i_L) \leq \exp\{-N^{\Omega(1)}\}$ , and since  $\mathbb{P}(\mathcal{E}_i) = N^{-O(1)}$  we have  $\mathbb{P}(T = i \mid \mathcal{E}_{i-1}) \leq \exp\{-N^{\Omega(1)}\}$  for every choice of  $i, u_1 \dots u_L$ . Thus, we can write

$$\mathbb{P}(\mathcal{E}_i \mid \mathcal{E}_{i-1}) = \frac{1}{Nq(1 \pm N^{-\epsilon\delta/2})} \pm \exp\{-N^{\Omega(1)}\} = \frac{(1 + o(1))}{Nq}.$$

If  $i_k < i < i_{k+1}$ , then  $\mathbb{P}(\mathcal{E}_i \mid \mathcal{E}_{i-1})$  is the probability that the set of vertices  $\{u_{k+1}, \dots, u_L\}$  all stay open and do not obtain an edge and we do not trigger the stopping time  $T$ . That is, we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}_i \mid \mathcal{E}_{i-1}) &= 1 - \frac{1}{Q(i-1)} \left[ \sum_{m=k+1}^L d_2(u_w) + O\left(C_{2,2 \rightarrow 1} + \sum_{m \geq 2} D_{m \uparrow m+1}\right) \right] \\ &= 1 - \frac{(L+1-k)(r-1)D^{\frac{1}{r-1}}t^{r-2}q \cdot (1 \pm N^{-\epsilon\delta/2})}{Nq \cdot (1 \pm N^{-\epsilon\delta/2})} \\ &= 1 - \frac{(L+1-k)(r-1)D^{\frac{1}{r-1}}t^{r-2}}{N} \cdot (1 \pm O(N^{-\epsilon\delta/2})) \end{aligned}$$

Thus, setting  $i_0 = 0$  we have

$$\begin{aligned} \mathbb{P}(\mathcal{E}) &= \prod_{k=1}^L \left[ \prod_{i=i_{k-1}}^{i_k-2} 1 - \frac{(L+1-k)(r-1)D^{\frac{1}{r-1}}t^{r-2} \cdot (1 \pm O(N^{-\epsilon\delta/2}))}{N} \right] \frac{1 + o(1)}{Nq(t(i_k - 1))} \\ &= (1 + o(1)) \exp \left\{ -\frac{(r-1)D^{\frac{1}{r-1}}}{N} \sum_{k=1}^L (L+1-k) \sum_{i=i_{k-1}}^{i_k-2} t^{r-2} \right\} \prod_{k=1}^L \frac{1}{Nq(t(i_k - 1))} \\ &= (1 + o(1)) \exp \left\{ -\frac{D^{\frac{1}{r-1}}}{N} \sum_{k=1}^L \sum_{i=0}^{i_k-2} (r-1)t^{r-2} \right\} \prod_{k=1}^L \frac{1}{Nq(t(i_k - 1))} \\ &= (1 + o(1)) \exp \left\{ -\sum_{k=1}^L t(i_k - 1)^{r-1} + O\left(\frac{t(i_{\max})^{r-2}D^{\frac{1}{r-1}}}{N}\right) \right\} \prod_{k=1}^L \frac{1}{Nq(t(i_k - 1))} \\ &= (1 + o(1)) \frac{1}{N^L} \end{aligned}$$

where on the last line we have used  $q(t) = \exp\{-t^{r-1}\}$ .

We complete the proof by summing over all possible choices of the indices  $i_k$ .  $\square$

Now by linearity of expectation, we have  $E[X_{\mathcal{G}}] = |\mathcal{G}|p^s \cdot (1 + o(1))$ . Now we will do a second moment calculation to show that  $X_{\mathcal{G}}$  is concentrated around its mean. It suffices to show that  $E[X_{\mathcal{G}}^2] \leq E[X_{\mathcal{G}}]^2 \cdot (1 + o(1))$ .

We have

$$E[X_{\mathcal{G}}^2] = \sum_{e, e' \in \mathcal{G}} \mathbb{P}(e \cup e' \subseteq I(i)).$$

Thus, by an application of the Lemma, we have

$$\begin{aligned} E[X_{\mathcal{G}}^2] &= \sum_{e \in \mathcal{G}} \sum_{a=0}^s |\{e' \in \mathcal{G} : |e \cap e'| = a\}| p^{2s-a} (1 + o(1)) \\ &\leq |\mathcal{G}|^2 p^{2s} (1 + o(1)) + O\left(|\mathcal{G}| \sum_{a=1}^s \Delta_a(\mathcal{G}) p^{2s-a}\right) \\ &= (1 + o(1)) E[X_{\mathcal{G}}]^2. \end{aligned}$$

To help see the second line, note that the  $1 + o(1)$  from the line above is unaffected by the sums since it is uniform over the terms. To see the last line note that  $|\mathcal{G}| \Delta_a(\mathcal{G}) p^{2s-a} = o(|\mathcal{G}|^2 p^{2s})$  for each  $a$ .

**Acknowledgment.** The authors thank Jacob Fox, Dhruv Mubayi, Mike Picolletti and Jozsef Balogh for helpful conversations. We also thank the anonymous referees for many helpful comments.

## References

- [1] M. Ajtai, J. Komlós, J. Pintz, J. Spencer, E. Szemerédi, Extremal uncrowded hypergraphs. *J. Combin. Theory Ser. A* **32** (1982) 321–335.
- [2] P. Bennett, The random greedy algorithm for sum-free subsets of  $\mathbb{Z}_{2n}$ , *arXiv:1502.01644*
- [3] T. Bohman, The triangle-free process, *Advances in Mathematics* **221** (2009) 1653–1677.
- [4] T. Bohman, A. Frieze, E. Lubetzky, A note on the random greedy triangle packing algorithm. *Journal of Combinatorics* **1** (2010), 477–488.
- [5] T. Bohman, A. Frieze, E. Lubetzky, Random triangle removal, *arXiv:1203.4223*
- [6] T. Bohman and P. Keevash, The early evolution of the H-free process, *Invent. Math.* **181** (2010), 291–336.
- [7] T. Bohman and P. Keevash, Dynamic concentration of the triangle-free process. *arXiv:1302.5963*
- [8] T. Bohman and M. Picolletti, Evolution of SIR epidemics on random graphs with a fixed degree sequence, *Random Structures and Algorithms* **41** (2012), 179–214.

- [9] J. Cooper and D. Mubayi, Sparse hypergraphs with low independence numbers, *Combinatorica*, to appear.
- [10] D. Conlon, J. Fox, Y. Zhao, A relative Szemerdi theorem. *arXiv:1305.5440*.
- [11] R. Duke, H. Lefmann, V. Rödl, On uncrowded hypergraphs. *Random Structures Algorithms* **6** (1995) 209–212.
- [12] G. Fiz Pontiveros, S. Griffiths and R. Morris, The triangle-free process and  $R(3,k)$ . *arXiv:1302.6279*
- [13] D. A. Freedman, On tail probabilities for martingales, *Ann. Probability* **3** (1975), 100–118.
- [14] S. Gerke and T. Makai, No dense subgraphs appear in the triangle-free graph process, *Electron. J. Combin.* **18** (2011), R168.
- [15] W. T. Gowers, Decompositions, approximate structure, transference, and the Hahn-Banach theorem. *arXiv:0811.3103*.
- [16] B. Green, personal communication.
- [17] D. Gunderson, V. Rödl, and A. Sidorenko, Extremal problems for sets forming Boolean algebras and complete partite hypergraphs, *Journal of Combinatorial Theory Ser. A* **88** (1999) 342–367.
- [18] W. Hoeffding, Probability inequalities for sums of bounded variables. *Journal of the American Statistical Association* **58** (1963), 13–30.
- [19] J.H. Kim, The Ramsey number  $R(3,t)$  has order of magnitude  $t^2/\log t$ , *Random Structures Algorithms* **7** (1995), 173–207.
- [20] D. Mubayi, Some exact results and new asymptotics for hypergraph Turán numbers, *Combinatorics, Probability and Computing* **11** (2002), 299–309.
- [21] M. Picollelli, The diamond-free process, *Random Structures and Algorithms* **45** (2014) 513–551.
- [22] M. Picollelli, The final size of the  $C_4$ -free process, *Combin. Probab. Comput.* **20** (2011), 939–955.
- [23] M. Picollelli, The final size of the  $C_l$ -free process, *SIAM Journal on Discrete Mathematics* **28** (2014) 1276–1305.
- [24] A. Ruciński and N. Wormald, Random graph processes with degree restrictions, *Combin. Probab. Comput.* **1** (1992), 169–180.
- [25] A. Telcs, N. Wormald and S. Zhou, Hamiltonicity of random graphs produced by 2-processes, *Random Structures and Algorithms* **31** (2007), 450–481.
- [26] L. Warnke, Dense subgraphs in the  $H$ -free process *Disc. Math.* **333** (2011), 2703–2707.

- [27] L. Warnke, The  $C_\ell$ -free process, *Random Structures Algorithms*, **44** (2014) 490–526.
- [28] L. Warnke, When does the  $K_4$ -free process stop? *Random Structures Algorithms*, **44** (2014) 355–397.
- [29] G. Wroblewski, Lower bounds for the size of random maximal H-free graphs. *Electronic J. Combin.* **16**, 2009, R4.
- [30] G. Wroblewski, The  $K_4$ -free process, arXiv:1008.4044.
- [31] G. Wroblewski, Triangle-free subgraphs in the triangle-free process, *Random Structures Algorithms* **39** (2011), 539–543.
- [32] N. Wormald, The differential equation method for random graph processes and greedy algorithms, in *Lectures on Approximation and Randomized Algorithms* (M. Karonski and H.J. Prömel, eds), pp. 73–155. PWN, Warsaw, 1999.