Testing the independence number of hypergraphs

Michael Langberg ⋆

Abstract. A k-uniform hypergraph G of size n is said to be ε-far from having an independent set of size ρn if one must remove at least εn^k edges of G in order for the remaining hypergraph to have an independent set of size ρn. In this work, we present a natural *property testing* algorithm that distinguishes between hypergraphs which have an independent set of size $\geq \rho n$ and hypergraphs which are ε-far from having an independent set of size ρn. Our algorithm is natural in the sense that we sample $\simeq c(k)\frac{\rho^{2k}}{\varepsilon^3}$ random vertices of G, and according to the independence number of the hypergraph induced by this sample, we distinguish between the two cases above. Here $c(k)$ depends on k alone (*e.g.* the sample size is independent of n). To the best of our knowledge, property testing of the independence number of hypergraphs has not been addressed in the past.

1 Introduction

A k-uniform hypergraph is a hypergraph $G = (V, E)$ in which each (hyper) edge is of size exactly k. An independent set I in G, is a subset of vertices that do not include any edges (*i.e.* there does not exist an edge $\{v_1, \ldots, v_k\} \in E$ for which $v_i \in I$ for all $i \in \{1, \ldots, k\}$). The size of the maximum independent set in G is denoted by $\alpha(G)$, and referred to as the independence number of G. Consider a k-uniform hypergraph G of (vertex) size n, with a maximum independent set I of size $\alpha(G)$. Let H be a random subgraph of G of size s (*i.e.* H is the subgraph induced by a random subset of vertices of G of size s). In this work we study the independence number of H.

For example, if it is the case that $\alpha(G) \geq \rho n$ for some parameter ρ, it is not hard to verify that with constant probability $\alpha(H)$ will be at least ρs. This follows from the fact that the expected size of $H \cap I$ is at least ρs. However, if $\alpha(G) < \rho n$, and this is our only assumption on G, we cannot hope for a positive bound on the probability that $\alpha(H) < \rho s$ (unless s is equal to n).

Thus, in the latter case, we strengthen our assumption on G to hypergraphs G which not only satisfy $\alpha(G) < \rho n$ but are also *far* from having an independent set of size ρn (we defer defining the exact notion of "far" until later in this discussion). That is, given a hypergraph G which is *far* from having an independent set of size ρn, we study the minimal value of s for which (with high probability) a random subgraph of size s does not have an independent set of size ρs. This

⋆ Department of Computer Science, California Institute of Technology, Pasadena, California 91125. E-mail: `mikel@cs.caltech.edu`. Part of this work was done while visiting the School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel.

question (and many other closely related ones) have been studied in (2-uniform hyper) graphs in [GGR98] under the title of *property testing*.

Property testing: Let $\mathcal{C}$ be a class of objects, and $\mathcal{P}$ a property of objects from $\mathcal{C}$. Property testing addresses the problem of distinguishing between elements $c \in \mathcal{C}$ which have the property $\mathcal{P}$ and elements that are *far* from having the property $\mathcal{P}$. The aim is to construct efficient (randomized) distinguishing algorithms that sample the given element c in relatively few places. The notion of property testing was first presented by Rubinfeld and Sudan in [RS96] where the testing of algebraic properties of functions was addressed. Goldreich, Goldwasser, and Ron [GGR98] later initiated the study of combinatorial objects in the context of property testing. In their work they studied (2-uniform) graphs and considered several fundamental combinatorial graph properties related to the independence number, chromatic number, size of maximum cut, and size of the maximum bisection of these graphs. Since, many papers have addressed the notion of property testing, both in the context of functions and in the combinatorial setting (*e.g.* see surveys [Gol98,Ron01,Fis01]).

Property testing of hypergraphs has also been studied in the past. Czumaj and Sohler [CS01] initiated this line of study when analyzing the property of being ℓ colorable. Colorability, and other properties of k-uniform hypergraphs (that can be phrased as a Max-k-CNF formula) were also studied in [AdlVKK02,AS03]. In this work, we consider testing the independence number of hypergraphs. To the best of our knowledge, this property has not been addressed in the past (in the context of hypergraphs).

Testing the independence number: Goldreich, Goldwasser, and Ron [GGR98] study property testing of the independence number of (2-uniform) graphs. In [GGR98] a graph G of size n is said to be ε-far from having an independent set of size ρn if any set of size ρn in G has at least εn^2 induced edges. It was shown in [GGR98] that if G is ε-far from having an independent set of size ρn then with high probability a random subgraph of size $s = \frac{c \log(1/\varepsilon)\rho}{\varepsilon^4}$, for a sufficiently large constant c, does not have an independent set of size ρs. The sample size s was later improved in [FLS02] to $c\frac{\rho^4}{\varepsilon^3}\log\left(\frac{\rho}{\varepsilon}\right)$ (again c is a sufficiently large constant). It is not hard to verify that this implies a (two-sided error) property testing algorithm for the independence number of G. Namely, given a graph G, one may sample a random subgraph H of G of size s, and exhaustively compute $\alpha(H)$. On one hand, if G happens to have an independent set of size ρn, then with some constant probability p_1 the independence number of H will be at least ρs (recall that the expected value of $\alpha(H)$ is at least ρs). On the other hand, if G is ε-far from having an independent set of size ρn then, as mentioned above, with at most some small probability p_2 (which is set to be smaller than p_1) it is the case that $\alpha(H) \geq \rho s$.

Our results: In this work, we address the study of hypergraphs in the above context. Namely, given a k-uniform hypergraph $G = (V, E)$ of size n, we show that if one must remove at least εn^k edges of G in order for the remaining hypergraph to have an independent set of size ρn (*i.e.* G is ε-far from having an independent set of size ρn) then a random subgraph H of G of size $s \simeq c(k)\frac{\rho^{2k}}{\varepsilon^3}$ satisfies $\alpha(H) < \rho s$ with high probability. Here $c(k)$ depends on k alone, which implies that the sample size s is independent of n. We also show a lower bound for the size of s of value $c'(k)\frac{\rho^{2k-1}}{\varepsilon^2}$ (where $c'(k)$ is yet an additional function which depends on k alone).

Definition 1 *Let A be a subset of V. Define $E(A)$ to be the set of edges in the hypergraph induced by A.*

Definition 2 *Let $\rho < 1$. A k-uniform hypergraph $G = (V, E)$ is said to be $\langle \rho, \varepsilon \rangle$-connected iff every subset A of V of size ρn satisfies $|E(A)| \geq \varepsilon n^k$ (*i.e. *the number of edges in the subgraph induced by A is greater than εn^k).*

Theorem 3. *Let G be a k-uniform hypergraph. Let H be a random sample of G of size $s \geq c2^k k! \frac{\rho^{2k}}{\varepsilon^3} \log\left(\frac{\rho}{\varepsilon}\right)$ for a large constant c.*

1. *If G has an independent set of size ρn, then with probability $\geq 1/4$ the subgraph H will have an independent set of size ρs.*
2. *If G is $\langle \rho, \varepsilon \rangle$-connected then with probability $\leq 1/20$ the subgraph H will have an independent set of size ρs.*

A few remarks are in place. First of all, in the above theorem we consider only hypergraphs which are k-uniform. Theorem 3 can be extended to hypergraphs in which any edge is of size at most k. This follows from the fact that any sufficiently large hypergraph which is $\langle \rho, \varepsilon \rangle$-connected with edges of size k and smaller contains a subgraph (on the same vertex set) which is k-uniform and $\langle \rho, \varepsilon/2 \rangle$-connected. Secondly, throughout this work we analyze the properties of random subsets H which are assumed to be *small*. Namely, we assume that the value of s (as stated in Theorem 3) and the parameters ρ, ε and n satisfy (a) $s < c\sqrt{n}$ and (b) $s < c\rho n$ for a sufficiently small constant c. These assumptions typically hold in the context of property testing. Finally, the constants 1/4, and 1/20 presented above are not tight (we have not made an attempt to find the tightest possible constants).

Proof techniques: Given a k-uniform hypergraph G with vertex size n which is $\langle \rho, \varepsilon \rangle$-connected, the bulk of our work addresses the study of the minimal value of s for which a random subgraph H of G satisfies $\alpha(H) < \rho s$ with high probability. This is done by analyzing the probability that a random subset H of G of size s satisfies $\alpha(H) < \rho s$ (as a function of s, ρ and ε). Our proof structure is strongly based on that appearing in (the journal version) of [FLS02], and proceeds as follows. Given a sample size s, we start by bounding the probability that a random subset R of G of size $\ell \geq \rho s$ is an independent set. As any subset R of H is random in G, one may bound the probability that $\alpha(H) \geq \rho s$ using the standard union bound on all subsets R of H of size $\geq \rho s$. We analyze this naive strategy and show that it only suffices to bound the probability of a slightly stronger condition than the condition $\alpha(H) \geq \rho s$. Namely, using this scheme, we bound the probability for which $\alpha(H) \geq 3\rho s$ (instead of exactly ρs). The naive strategy is then enhanced in order to bound the probability that $\alpha(H) > \delta\rho s$ as a function of δ for any $\delta > 1$. This now suffices to bound the probability that $\alpha(H) \geq \rho s$ (using a few additional ideas).

Applying the proof technique of [FLS02] to the case of k-uniform hypergraphs involves several difficulties. Roughly speaking, these are overcome by considering for each vertex v in the given graph G, a weighted set system of all subsets of V which *share* an edge with v. This set system plays the role of the standard "set of neighbors" considered when studying (two-uniform) graphs. Specifically, for all vertices v we consider the set system $\mathcal{A}_v$ consisting of sets $\alpha = \{u_1, \ldots, u_i\}$ for which there exists an edge e in G which includes both the vertex v and the set α. The weight applied to each subset is fixed to be a function of the subset size. Ideas along this line have been used in the past in the study of hypergraphs (*e.g.* [AS03]).

Organization: The remainder of the paper is organized as follows. In Section 2 we present the proof of Theorem 3. In Section 3 we present our lower bounds on the sample size s. As mentioned above, our proof uses many of the arguments presented in [FLS02]. During our presentation, we repeat quite a few of these arguments, in aim to make the paper comprehensible. Due to space limitations, some of our results appear without detailed proof. Most of the details can be found in an extended version of this paper [Lan03].

2 Proof of Theorem 3

In this section, we prove Theorem 3. The proof of Theorem 3 (1) is straightforward and (due to space limitations) is omitted. The proof of Theorem 3 (2) follows.

2.1 The Naive scheme

Let $G = (V, E)$ be a $\langle\rho, \varepsilon\rangle$-connected k-uniform hypergraph. In this section we study the probability that a random subset R of V of size ℓ is an independent set. We then use this result to bound the probability that a random subset H of G of size s has a large independent set.

We would like to bound (from above) the probability that R induces an independent set. Let $\{r_1, \ldots, r_\ell\}$ be the vertices of R. Consider choosing the vertices of R one by one, such that at each step the random subset chosen so far is $R_i = \{r_1, \ldots, r_i\}$ and the vertex r_{i+1} is chosen from $V \setminus R_i$. Assume that at some stage R_i is an independent set. We would like to show (with high probability) that after adding the remaining vertices $\{r_{i+1}, \ldots, r_\ell\}$ to R_i, the final set R will not be an independent set.

The vertices in $V \setminus R_i$ that *cannot* be added to R_i are exactly the vertices v that share an edge with some $k-1$ vertices in R_i (namely, these vertices will result in a set R_{i+1} that is no longer independent). Let $N(R_i) = N_i$ be the set of such vertices in $V \setminus R_i$, and let $I(R_i) = I_i$ be $V \setminus N_i$. Consider the next random vertex $r_{i+1} \in R$. If r_{i+1} is chosen from N_i then it cannot be added to R_i, and we view this round as a success regarding the set R_i. Otherwise, r_{i+1} happens to be in I_i and can be added to R_i. But if the addition of r_{i+1} to R_i happens to add many vertices to N_i, we also view this round as a successful round regarding R_i.

Motivated by the discussion above, we continue with the following definitions. For each subset R_i we define the following weighted set systems of subsets of V. Let $RES(R_i, 1)$ (for *restrict*) be the set of singletons $\{v\}$ that share an edge with vertices in R_i. Namely, $\{v\} \in RES(R_i, 1)$ iff there exists vertices $\{w_1, \ldots, w_{k-1}\}$ in R_i such that there is an edge $\{v, w_1, \ldots, w_{k-1}\}$ in E. $RES(R_i, 1)$ is exactly the set N_i defined above. Define the weight of each element in $RES(R_i, 1)$ as n^{k-2}. Let $RES(R_i, 2)$ be the set of *pairs* of vertices v_1, v_2 which (together) share an edge with vertices in R_i. Namely, $\{v_1, v_2\} \in RES(R_i, 2)$ iff there exists vertices $\{w_1, \ldots, w_{k-2}\}$ in R_i such that there is an edge $\{v_1, v_2, w_1, \ldots, w_{k-2}\}$ in E. Define the weight of each element in $RES(R_i, 2)$ as n^{k-3}. Similarly for each $j \in \{1, \ldots, k-1\}$ let $RES(R_i, j)$ be the set of subsets $\{v_1, \ldots, v_j\}$ of V of size j that share an edge with vertices in R_i. Define the weight of each element in $RES(R_i, j)$ as n^{k-j-1}. Finally let $RES_i = RES(R_i)$ be the union of the sets $RES(R_i, j)$ where $j \in \{1, \ldots, k-1\}$. Let $\|RES(R_i, j)\|$ ($\|RES_i\|$) denote the weight of elements in $RES(R_i, j)$ (RES_i). Notice that $\|RES(R_i, j)\| \leq \binom{n}{j} n^{k-j-1} \leq n^{k-1}$ and that $\|RES_i\| \leq \sum_{j=1}^{k-1} \binom{n}{j} n^{k-j-1} \leq kn^{k-1}$.

Definition 4 *Let the normalized degree w.r.t. R_i of a vertex $v \in V$ be the amount on which v restricts upon R_i:*

$$d_v(R_i) = \|RES(R_i \cup \{v\})\| - \|RES(R_i)\| = \|RES(R_i \cup \{v\}) \setminus RES(R_i)\|.$$

In the above notice that $RES(R_i) \subseteq RES(R_i \cup \{v\})$. We call a vertex v in V heavy with respect to R_i (or R_i-heavy for short) if $d_v(R_i) \geq \frac{1}{2^k (k-3)!} \frac{\varepsilon}{\rho} n^{k-1}$.

Each subset R_i of V now defines the following partition (LI_i, HI_i, N_i) of V and the set RES_i. Let RES_i and N_i be as defined as above. Let $I_i = V \setminus N_i$. I_i is now partitioned into two parts: vertices in I_i with *low* normalized degree (w.r.t. R_i), denoted as the set LI_i, and vertices with *high* normalized degree, denoted as HI_i. Namely LI_i is defined to be the ρn vertices of I_i with minimal normalized degree and HI_i is defined to be the remaining vertices of I_i. Ties are broken arbitrarily or in *favor* of vertices in R_i (namely, vertices in R_i are placed in LI_i before other vertices of identical degree). If it is the case that $|I_i| \leq \rho n$ then LI_i is defined to be I_i, and HI_i is defined to be empty.

We define the partition corresponding to $R_0 = \phi$ as (LI_0, HI_0, N_0), where LI_0 are the ρn vertices of G of minimal normalized degree, HI_0 are the remaining vertices of G, and N_0 is empty. RES_0 is also defined to be empty.

Notice, using this notation, that the subset R_i is an independent set iff $R_i \subseteq I_i$. Moreover, in this case $R_i \subseteq LI_i$ (all vertices of R_i have normalized degree 0). Furthermore, each vertex r_i in an independent set $R = R_\ell = \{r_1, \ldots, r_\ell\}$ satisfies $r_i \in I_{i-1}$.

We are now ready to bound the probability that a random subset $R = \{r_1, \dots, r_\ell\}$ of G is independent. Let $R_i = \{r_1, \dots, r_i\}$, and let (LI_i, HI_i, N_i) be the corresponding partition of V defined by R_i. Consider the case in which R is an independent set. As mentioned above, this happens iff for every i the vertex r_i is chosen to be in $I_{i-1} = LI_{i-1} \cup HI_{i-1}$. We would like to show that this happens with small probability (if ℓ is large enough).

Consider the set RES_i as we proceed in the choice of vertices in R. Initially, the subset RES_0 is of weight 0, and it gets larger and larger as we proceed in the choice of vertices in R. Each vertex in $r_i \in HI_{i-1}$ increases the weight of RES_{i-1} substantially, while each vertex in LI_{i-1} may only slightly change the weight of RES_{i-1}. In the following, we show that there cannot be many vertices $r_i \in R$ that happen to fall into HI_{i-1}. We thus turn to consider vertices r_i that fall in LI_{i-1} (there are almost ℓ such vertices). The size of LI_i is bounded by ρn. Hence, the probability that $r_i \in LI_i$ is bounded by ρ (by our definitions $R_{i-1} \subseteq LI_{i-1}$ and the vertex r_i is random in $V \setminus R_{i-1}$). This implies that the probability that R is an independent set is roughly bounded by ρ^ℓ. Details follow.

Lemma 1. *Let G be a $\langle \rho, \varepsilon \rangle$-connected hypergraph. Let $R = \{r_1, \dots, r_\ell\}$ be a set in G. The number of vertices r_i which satisfy $r_i \in HI_{i-1}$ is bounded by $t = \frac{2^k k! \rho}{\varepsilon}$.*

Proof. We start with the following claim which is the main technical contribution of this work.

Claim 5 *Let R_i be as defined above, and let (LI_i, HI_i, N_i) and RES_i be its corresponding partition and set system. Let $I_i = LI_i \cup HI_i$. If G is $\langle \rho, \varepsilon \rangle$-connected then, every vertex in HI_i is R_i-heavy.*

Proof. Assume that $LI_i = \rho n$ (otherwise HI_i is empty and the claim holds). By our assumptions on G we have that LI_i induces at least εn^k edges. Let $E(LI_i)$ be the set of these edges, and let m be the size of $E(LI_i)$. To simplify our notation, let $LI = LI_i$, $HI = HI_i$, $N = N_i$ and $R = R_i$. Now that the index i is free, we use it in the following new context. For $i \in \{1, \dots, k-1\}$ consider the following sequence of weighted i-uniform hypergraphs $H_i = (LI, E_i)$ which all have vertex set LI. The edge set E_i is defined to be all subsets of size i of edges in $E(LI)$ (here and throughout our work we consider a (hyper) edge of size k as a subset of vertices of size k). For example, each edge in $E(LI)$ induces $\binom{k}{2}$ edges in H_2, $\binom{k}{3}$ edges in H_3, and so on. The weight of an edge e_i in H_i is equal to the number of edges e in $E(LI)$ which satisfy $e_i \subseteq e$.

Recall the sets $RES(R, j)$ and $RES(R)$. Our goal is to prove the existence of a vertex $v \in LI$ which is R-heavy (this will imply our assertion). Namely a vertex v which when added to R will significantly increase the weight of $RES(R)$. Formally we are looking for a vertex $v \in LI$ which satisfies $\|RES(R \cup \{v\})\| - \|RES(R)\| \geq \frac{1}{2^k (k-3)!} \frac{\varepsilon}{\rho} n^{k-1}$. Notice that the weight of a subset of vertices with respect to $RES(R)$ may (and usually does) differ from its weight with respect to H_i. To avoid confusion we denote the weight of a subset e with respect to $RES(R)$ as $w_R(e)$. Recall that every (partial) edge e of size i in $RES(R)$ has weight $w_R(e) = n^{k-i-1}$. The weight of an edge e in H_i will be marked as $w_{H_i}(e)$.

For each i we now consider the weight (with respect to H_i) of the edges in H_i which are in $RES(R)$. Denote this weight as res_i. Notice that $res_1 = 0$ (by our construction $LI \cap N = \phi$, which implies that singletons $\{v\} \subseteq LI$ are not in $RES(R)$). We consider the following cases.

Case 1: We start by assuming that $res_{k-1} \leq (k-1)m$. As the total weight of edges in H_{k-1} is km, we conclude that the weight of edges in H_{k-1} that are not in $RES(R)$ is at least m. Let v be a vertex in LI. We say that an edge e in H_{k-1} is a v-edge if there exists an edge in $E(LI)$ consisting of the union of e and $\{v\}$. Notice that an edge e in H_{k-1} of weight $w_{H_{k-1}}(e)$ is actually a v edge for $w_{H_{k-1}}(e)$ distinct vertices v. As each edge in H_{k-1} is a v-edge for some vertex v in LI, we have the existence of a vertex v in LI with at least $\frac{m}{\rho n}$ (distinct) v-edges that are not in $RES(R)$. The weight of each such v-edge in $RES(R)$ is 1. Furthermore, each such v-edge appears in $RES(R \cup \{v\})$ implying that

$$\|RES(R \cup \{v\})\| - \|RES(R)\| \geq \frac{m}{\rho n} \geq \frac{\varepsilon}{\rho} n^{k-1}$$

which in turn implies the v is R-heavy.

Cases 2 to $k-1$: Starting at $i = k-1$ and iteratively continuing until $i = 2$ consider the following cases. From the previous step, we may assume that $res_i > \frac{2m}{2^{k-i}} \frac{(k-1)}{(k-i)!}$. We now also assume that $res_{i-1} \leq \frac{m}{2^{k-i}} \frac{(k-1)}{(k-i+1)!}$. Consider an edge $e = (v_1, \ldots, v_i)$ in H_i of weight $w_{H_i}(e)$ which contributes to res_i. Each such edge induces i edges in H_{i-1}: $\{e'_1, \ldots, e'_i\}$ (each edge obtained by removing one vertex from e). We are interested in bounding (by below) the weight of edges e as above with corresponding edges e'_j which are not in $RES(R)$ (here $j \in \{1, \ldots, i\}$).

By our construction, the weight of any edge e' in H_{i-1} equals the number of edges in G that include e'. It is not hard to verify that this equals $\frac{1}{k-i+1}$ times the weight of edges in H_i which include e'. We conclude that the weight of edges e in H_i which are in $RES(R)$ with some corresponding edge e' in H_{i-1} which is also in $RES(R)$ is bounded by $res_{i-1}(k-i+1)$. This leaves us with

$$res_i - (k-i+1)res_{i-1} > \frac{m}{2^{k-i}} \frac{(k-1)}{(k-i)!}$$

edges e which contribute to res_i for which e'_j for $j \in \{1, \ldots, i\}$ are not in $RES(R)$.

We now conclude the existence of a vertex $v \in LI$ which is adjacent to (*i.e.* is included in) at least the weight of $\frac{im}{2^{k-i}\rho n} \frac{(k-1)}{(k-i)!}$ edges in H_i which appear in $RES(R)$ such that their corresponding edges in H_{i-1} are not in $RES(R)$. The weight of each edge in H_i is bounded by $\binom{\rho n}{k-i} \leq (\rho n)^{k-i}$. Thus there are at least $\frac{im}{2^{k-i}(\rho n)^{k-i+1}} \frac{(k-1)}{(k-i)!}$ distinct edges in H_i adjacent to v with corresponding edges in H_{i-1} that are not in $RES(R)$.

We will now show that v is R-heavy. Consider one of the distinct edges $e = (v, v_1, \ldots, v_{i-1})$ as discussed above. By our assumption the set $\{v_1, \ldots, v_{i-1}\}$

is not in $RES(R)$. The edge e appears in $RES(R)$ implying that there exists vertices $\{w_1, \ldots, w_{k-i}\}$ (in R) such that $(v, v_1, \ldots, v_{i-1}, w_1, \ldots, w_{k-i})$ is an edge in the original hypergraph G. This in turn implies that $\{v_1, \ldots, v_{i-1}\}$ will be included in $RES(R \cup \{v\})$. As each set of size $i-1$ in $RES(R)$ has weight $w_R = n^{k-i}$, and there are at least $\frac{im}{2^{k-i}(\rho n)^{k-i+1}} \frac{(k-1)}{(k-i)!}$ distinct edges of interest adjacent to v, we conclude that

$$\|RES(R \cup \{v\})\| - \|RES(R)\| \geq \frac{1}{2^{k-i}} \frac{(k-1)}{(k-i)!} \frac{\varepsilon}{\rho} n^{k-1}.$$

Now to prove our lemma, consider the subsets $R_i = \{r_1, \ldots, r_i\}$ and their corresponding partitions (LI_i, HI_i, N_i). Let $I_i = LI_i \cup HI_i$. We would like to bound the number of vertices r_i that are in HI_{i-1}. Consider a vertex r_i in HI_i. By Claim 5, its normalized degree w.r.t. R_i is at least $\frac{1}{2^k (k-3)!} \frac{\varepsilon}{\rho} n^{k-1}$. $RES(\phi)$ is initially empty, and for any i, $RES(R_i)$ is of weight at most $\sum_{i=2}^{k-1} \binom{n}{i} n^{k-i-1} \leq kn^{k-1}$. Each vertex $r_i \in HI_{i-1}$ increases $\|RES(R_{i-1})\|$ by at least $\frac{1}{2^k (k-3)!} \frac{\varepsilon}{\rho} n^{k-1}$. We conclude that there are at most $\frac{2^k k! \rho}{\varepsilon}$ vertices r_i in R which are in HI_{i-1}.

Let δ be a large constant. "Plugging" Lemma 1 into the proof appearing in (the journal version of) [FLS02], we are now able to bound the probability that a random subset H of G of size s has an independent set of size $> \delta\rho s$. This is done in two steps, first we bound the probability that a random subset R of G is independent. Then, using the union bound, we obtain the result mentioned for H. In the upcoming Section 2.2 we present the refined proof technique of [FLS02] (which *gets rid* of the parameter δ) applied to the case of hypergraphs. Namely, we will bound the probability that a random subset H of G of size s has an independent set of size $\geq \rho s$.

Lemma 2 ([FLS02]). *Let G be a $\langle \rho, \varepsilon \rangle$-connected hypergraph. Let t be as in Lemma 1. Let $\ell \geq 2t$. The probability that ℓ random vertices of G induce an independent set is at most*

$$\rho^\ell \left(\frac{e\ell}{t\rho} \right)^t$$

Corollary 6 ([FLS02]) *Let G be a $\langle \rho, \varepsilon \rangle$-connected hypergraph. Let t be as in Lemma 1. Let H be a random sample of G of size s. Let $\delta > e$, and let c be a sufficiently large constant. If $s \geq ct \frac{\log(1/\rho)}{\rho}$ then the probability that $\alpha(H) > \delta\rho s$ is at most $\left(\frac{e}{\delta}\right)^{\Omega(\delta\rho s)}$.*

2.2 An enhanced analysis

Recall the proof technique from Section 2.1. We started by analyzing the probability that a subset R of H of size ℓ is an independent set. Afterwards we bounded the probability that $\alpha(H) > \delta\rho n$ (using the standard union bound on all subsets R of H of size greater than $\ell = \delta\rho n$). In this section we enhance the first part of this scheme by analyzing the probability that a subset R of H of size ℓ is a *maximum* independent set in H (rather than just an independent set of H). Then, as before, using the standard union bound on all large subsets R of H, we bound the probability that $\alpha(H) \geq \rho s$. We show that taking the maximality property of R into account will suffice to prove Theorem 3 (2).

Let $H = \{h_1, \ldots, h_s\}$ be s random vertices in G. We would like to analyze the probability that a given subset R of H of size ℓ is a *maximum* independent set. Recall (Section 2.1), that the probability that R is an independent set is bounded by approximately ρ^ℓ. An independent set R is a maximum independent set in H only if adding any other vertex in H to R will yield a set which is no longer independent. Let $R = R_\ell$ be an independent set, and let $(LI_\ell, HI_\ell, N_\ell)$ be the partition (as defined in Section 2.1) corresponding to R. Consider an additional random vertex h from H. The probability that $R \cup h$ is no longer an independent set is approximately $|N_\ell|/n$ (here we assume that $|R|$ is small compared to n). The probability that for every $h \in H \setminus R$ the subset $R \cup \{h\}$ is no longer independent is thus $\simeq (|N_\ell|/n)^{s-\ell}$. Hence, the probability that a given subset R of H of size ℓ is a *maximum* independent set is bounded by approximately $\rho^\ell (|N_\ell|/n)^{s-\ell}$. This value is substantially smaller than ρ^ℓ iff $|N_\ell|$ is substantially smaller than n. We conclude that it is in our favor to somehow ensure that $|N_\ell|$ is not too large. We do this in an artificial manner.

Let $R = \{r_1, \ldots, r_\ell\}$ be an independent set, let $R_i = \{r_1, \ldots, r_i\}$, and let (LI_i, HI_i, N_i) be the partition (as defined in Section 2.1) corresponding to R_i. Let RES_i be the set system corresponding to R_i. Roughly speaking, in Section 2.1, every time a vertex r_i was chosen, the set system RES_i and N_i was updated. If r_i was chosen in HI_{i-1}, then RES_{i-1} increased substantially and N_{i-1} potentially also grew substantially, and if r_i was chosen in LI_{i-1}, both RES_{i-1} and N_{i-1} only slightly changed. We would like to change the definition of the partition (LI_i, HI_i, N_i) and of RES_i corresponding to R_i as to ensure that N_i is always substantially smaller than n. This cannot be done unless we relax the definition of N_i. Recall that N_i was defined (in Section 2.1) to be the set of vertices v which share an edge with vertices in R_i. Specifically, N_i is the set of vertices v for which there exists vertices $\{w_1, \ldots, w_{k-1}\}$ in R_i such that there is an edge $\{v, w_1, \ldots, w_{k-1}\}$ in E. In this section N_i will only include a subset of these vertices (a subset which is substantially smaller than n). Namely, in our new definitions RES_{i-1} and N_{i-1} will be changed only if r_i was chosen in HI_{i-1}. In the case in which $r_i \in LI_{i-1} \cup N_{i-1}$, we do not change N_{i-1} at all. As we will see, such a definition will imply that $|N_i| \leq (1-\rho)s$, which will now suffice for our proof.

A new partition and set system: Let $H = \{h_1, \ldots, h_s\}$ be a subset of V. Let $R_i = \{r_1, \ldots, r_i\}$ be a subset of H of size i. In the previous section, the subset R_i defined a partition (LI_i, HI_i, N_i) of V and a set system RES_i. In this section, for each i we will define a subset $\hat{R}_i$ of R_i, and a new partition and set system. The new partition and set system corresponding to R_i will be defined similarly to those defined in the previous section with the exception that $\hat{R}_i$ will play the role that R_i played previously. The set $\hat{R}_i$, the new partition (LI_i, HI_i, N_i), and the set system RES_i are defined as follows (as before, let $I_i = LI_i \cup HI_i$).

1. Initially $R_0 = \hat{R}_0 = \phi$, LI_0 is the ρn vertices in V of minimal normalized degree w.r.t. $\hat{R}_i$, $HI_0 = V \setminus LI_0$, and $N_0 = \phi$. In the above, ties are broken by an assumed ordering on the vertices in V. $RES_0 = RES(\hat{R}_0)$ is defined to be empty.
2. Let $\hat{R}_i$, (LI_i, HI_i, N_i) and RES_i be the sets corresponding to R_i, let r_{i+1} be a new random vertex. Let $R_{i+1} = R_i \cup \{r_{i+1}\}$, we now define the sets $\hat{R}_{i+1}$, $(LI_{i+1}, HI_{i+1}, N_{i+1})$ and RES_{i+1}. Let $N(r_{i+1})$ be the set of vertices which share an edge with vertices in $\hat{R}_i \cup r_{i+1}$. We consider the following cases: (a) If $r_{i+1} \in LI_i$ or $r_{i+1} \in N_i$ then the sets corresponding to R_{i+1} will be exactly those corresponding to R_i. Namely, $\hat{R}_{i+1} = \hat{R}_i$, $LI_{i+1} = LI_i$, $HI_{i+1} = HI_i$, $N_{i+1} = N_i$, and RES_{i+1} will be defined as RES_i. (b) If $r_{i+1} \in HI_i$ then we consider two sub-cases: (b1) If $|N_i \cup N(r_{i+1})| \leq (1-\rho)n$, then $\hat{R}_{i+1} = \hat{R}_i \cup \{r_{i+1}\}$, $RES_{i+1} = RES(\hat{R}_{i+1})$, and LI_{i+1}, HI_{i+1}, N_{i+1}, are defined as in Section 2.1. Namely, $N_{i+1} = N_i \cup N(r_{i+1})$. I_{i+1} is defined to be $V \setminus N_{i+1}$. LI_{i+1} is defined to be the ρn vertices of I_{i+1} with minimal normalized degree w.r.t. $\hat{R}_{i+1}$. Finally, HI_{i+1} is defined to be the remaining vertices of I_{i+1}. Ties are broken by the assumed ordering on V. (b2) If $|N_i \cup N(r_{i+1})| > (1-\rho)n$, then let $\hat{N}(r_{i+1})$ be the first (according to the assumed ordering on V) $(1-\rho)n - |N_i|$ vertices in $N(r_{i+1})$ and set $N_{i+1} = N_i \cup \hat{N}(r_{i+1})$. Furthermore, set LI_{i+1} to be the remaining ρn vertices of G, and HI_i to be empty. Notice that in this case $|N_{i+1}|$ is of size exactly $(1-\rho)n$. Finally, let $\hat{R}_{i+1} = \hat{R}_i$ and $RES_{i+1} = RES_i$.

A few remarks are in place. First of all it is not hard to verify that the definition above implies $|N_i| \leq (1-\rho)n$ for all $i \in \{1, \ldots, \ell\}$. Secondly, due to the iterative definition of our new partition, the sets corresponding to the subsets R_i depend strongly on the specific ordering of the vertices in R_i. Namely, in contrast to the partitions (set systems) used in Section 2.1, a single subset R with two different orderings may yield two different partitions (set systems). For this reason, in the remainder of this section, we will assume that the vertices of H are chosen one by one. This will imply an ordering on H and on any subset R of H. The partitions we will study will correspond to these orderings only.

Finally, in Section 2.1, an (ordered) subset $R = \{r_1, \ldots, r_\ell\}$ was independent iff $\forall i \;\; r_i \in I_{i-1}$ (according to the definition of I_{i-1} appearing in Section 2.1). In this section, if R is independent then it still holds that $\forall i \;\; r_i \in I_{i-1}$. However, it may be the case that $\forall i \;\; r_i \in I_{i-1}$ but R is not an independent set. In the remainder of this section, we call ordered subsets R for which $\forall i \;\; r_i \in I_{i-1}$ *free* sets. We analyze the probability that a random ordered subset H of V of size s does not have any free sets R of size larger then ρs. This implies, that H does not include any independent sets of size ρs.

We now turn to prove Theorem 3 (2). Roughly speaking, we start by analyzing the probability that a random subset R is a free set. We then analyze the probability that a given subset R in H is a maximum free set. Finally, we use the union bound on all subsets R of H of size $\geq \rho s$ to obtain our results.

We start by stating the following lemma which is analogous to Lemma 1 from Section 2.1. The main difference between the lemma below (and its proof), and that of the previous section is in the definition of the partition (LI_i, HI_i, N_i), and the set system RES_i. Proof of the lemma is omitted.

Lemma 3. *Let G be a $\langle\rho,\varepsilon\rangle$-connected hypergraph. Let $R = \{r_1, \ldots, r_\ell\}$ be an ordered set in G of size ℓ. The number of vertices r_i which satisfy $r_i \in HI_{i-1}$ is bounded by $t = \frac{2^k k!\rho}{\varepsilon}$.*

We now address the probability that a random subset R of H is a maximum free set. We then use the union bound on all subsets R of H of size $\geq \rho s$ (and a few additional ideas) to obtain our main result. Due to space limitations, the proof of the following Lemma and Theorem is omitted. The interested reader can find their proofs in [Lan03] (or proofs of similar nature in [FLS02]).

Lemma 4. *Let G be a $\langle\rho,\varepsilon\rangle$-connected hypergraph. Let t be as in Lemma 3. Let $\ell \geq 2t$. Let H be an ordered random sample of G of size $s \geq \ell$. The probability that a given subset R of H of size ℓ is a maximum free set is at most*

$$\rho^\ell \left(\frac{e\ell}{t\rho}\right)^t (1-\rho)^{s-\ell}$$

Theorem 3 (2). *Let G be a $\langle\rho,\varepsilon\rangle$-connected hypergraph. Let t be as in Lemma 3. Let H be a random sample of G of size s. Let c be a sufficiently large constant. If $s \geq ct\frac{\rho^{2k-1}}{\varepsilon^2}\log\left(\frac{\rho}{\varepsilon}\right)$ then the probability that H has an independent set of size $\geq \rho s$ is at most $e^{-\Omega(t)}$.*

3 Lower bounds on the sample size s

Roughly speaking, Theorem 3 (2) states that given a $\langle\rho,\varepsilon\rangle$-connected hypergraph G, a random sample H of G of size s proportional to $2^k k!\frac{\rho^{2k}}{\varepsilon^3}$ (or larger) will not have an independent set of size ρs (with high probability). We now continue to study the minimal value of s for which $\alpha(H) < \rho s$ with high probability, and present a lower bound of $\frac{\rho^{2k-1}}{4(k!)^2\varepsilon^2}$ on the size of s.

Lemma 5. *Let n be a sufficiently large constant. Let $\rho > 0$ and $\varepsilon > 0$ satisfy (a) $\varepsilon << \frac{\rho^k}{2k!}$, (b) $\rho^{2k-1}/\varepsilon^2 << n$ and (c) $k^2 << \rho n$ (here $a << b$ if for a sufficiently small constant $\delta > 0$ it holds that $a < \delta b$). There exists a graph G on n vertices for which G is $\langle\rho,\varepsilon\rangle$-connected, and with probability $\geq 1/20$ a random subgraph H of size $s = \frac{\rho^{2k-1}}{4(k!)^2\varepsilon^2}$ will have an independent set of size ρs.*

Proof. Consider the k-uniform hypergraph $G = (V, E)$ in which (a) $|V| = n$, (b) V consists of two disjoint sets A and $V \setminus A$, where A is of size $(1 - \frac{2k!\varepsilon}{\rho^k})\rho n$, and (c) the edge set E of G consists of all subsets of V of size k except those included in A (namely, A is an independent set). On one hand, every subset of size ρn in G induces a subgraph with at least εn^k edges (implying that G is $\langle \rho, \varepsilon \rangle$-connected). On the other, let H be a random subset of V obtained by picking each vertex independently with probability $\frac{1}{n} \frac{\rho^{2k-1}}{4(k!)^2 \varepsilon^2}$. The expected size of H is $s = \frac{\rho^{2k-1}}{4(k!)^2 \varepsilon^2}$. In the following, we assume H is exactly of size s, minor modifications in the proof are needed if this assumption is not made. The set $H \cap A$ is an independent set in the subgraph induced by H. The expected size of $H \cap A$ is $(1 - \frac{2k!\varepsilon}{\rho^k})\rho s$. Let $N(0,1)$ denote a standard normal variable. It can be seen using the central limit theorem (for example [Fel66]) that for our choice of parameters, the probability that $|H \cap A|$ is greater than $\left(1 - \frac{2k!\varepsilon}{\rho^k}\right) \rho s + \sqrt{\rho s}$ is at least $\Pr\left[N(0,1) > 3/2\right] \geq \frac{1}{20}$. In such a case the size of $H \cap A$ will be greater than $(1 - \frac{2k!\varepsilon}{\rho^k})\rho s + \sqrt{\rho s} = \rho s$ for our value of s. Hence implying the lemma.

References

[AdlVKK02] N. Alon, W. Fernandez de la Vega, R. Kannan, and M. Karpinski. Random sampling and approximation of Max-CSP problems. *In Proceddings of STOC*, pages 232–239, 2002.

[AS03] N. Alon and A. Shapira. Testing satisfyability. *Journal of Algorithms*, 47:87–103, 2003.

[CS01] A. Czumaj and C. Sohler. Testing hypergraph coloring. *In Proceddings of ICALP*, pages 493–505, 2001.

[Fel66] W. Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, 1966.

[Fis01] E. Fisher. The art of uninformed decisions: A primer to property testing. *The Computational Complexity Column of The Bulletin of the European Association for Theoretical Computer Science*, 75:97–126, 2001.

[FLS02] U. Feige, M. Langberg, and G. Schechtman. Graphs with tiny vector chromatic numbers and huge chromatic numbers. *In proceedings of 43rd annual Symposium on Foundations of Computer Science*, pages 283–292, 2002. *To appear in SIAM Journal on Computing, manuscript availiable at http://www.cs.caltech.edu/∼mikel.*

[GGR98] O. Goldreich, S. Goldwasser, and D. Ron. Property testing and its connection to learning and approximation. *Journal of ACM*, 45(4):653–750, 1998.

[Gol98] O. Goldreich. Combinatorial property testing - a survey. *Randomization Methods in Algorithm Design (P. Pardalos, S. Rajasekaran and J. Rolim eds.), AMS-DIMACS*, pages 45–60, 1998.

[Lan03] M. Langberg. Testing the independence number of hypergraphs. *Electronic Colloquium on Computational Complexity (ECCC))*, TR03-076, 2003.

[Ron01] D. Ron. Property testing (a tutorial). *In Handbook of Randomized Computing (S. Rajasekaran, P. M. Pardalos, J. H. Reif and J. D. P. Rolim eds), Kluwer Press*, 2001.

[RS96] R. Rubinfeld and M. Sudan. Robust characterization of polynomials with applications to program testing. *SIAM Journal of Computing*, 25:252–271, 1996.