A Proof of Green's Conjecture Regarding the Removal Properties of Sets of Linear Equations[*]

Asaf Shapira [†]

Dedicated to the Memory of Oded Schramm

Abstract

A system of ℓ linear equations in p unknowns $Mx = b$ is said to have the *removal property* if every set $S \subseteq \{1, \ldots, n\}$ which contains $o(n^{p-\ell})$ solutions of $Mx = b$ can be turned into a set S' containing no solution of $Mx = b$, by the removal of $o(n)$ elements. Green [GAFA 2005] proved that a single homogenous linear equation always has the removal property, and conjectured that every set of homogenous linear equations has the removal property. In this paper we confirm Green's conjecture by showing that every set of linear equations (even non-homogenous) has the removal property.

We also discuss some applications of our result in theoretical computer science, and in particular, use it to resolve a conjecture of Bhattacharyya, Chen, Sudan and Xie [9] related to algorithms for testing properties of boolean functions.

MSC-2000 classification: 05C35, 11P99 ,68R99

1 Introduction

1.1 Background

The (triangle) removal lemma of Ruzsa and Szemerédi [30], which is by now a cornerstone result in combinatorics, states that a graph on n vertices that contains only $o(n^3)$ triangles can be made triangle free by the removal of only $o(n^2)$ edges. Or in other words, if a graph has asymptotically few triangles then it is asymptotically close to being triangle free. While the lemma was proved in [30] for triangles, an analogous result for any fixed graph can be obtained using the same proof

[*] A Preliminary version of this paper, under the title *"Green's Conjecture and Testing Linear-Invariant Properties"*, appeared in the Proc. of the 41^{st} ACM Symposium on Theory of Computing (STOC) 2009.

[†] School of Mathematics and College of Computing, Georgia Institute of Technology, 686 Cherry St., Atlanta, GA 30332. Email: asafico@math.gatech.edu. Research supported in part by NSF Grant DMS-0901355.

idea. Actually, the main tool for obtaining the removal lemma is Szemerédi's regularity lemma for graphs [36], another landmark result in combinatorics. The removal lemma has many applications in different areas like extremal graph theory, additive number theory and theoretical computer science. Perhaps its most well known application appears already in [30] where it is shown that an ingenious application of it gives a very short and elegant proof of Roth's Theorem [27], which states that if $S \subseteq [n] = \{1, \dots, n\}$ has no 3-term arithmetic progression then $|S| = o(n)$.

Recall that an r-uniform hypergraph $H = (V, E)$ has a set of vertices V and a set of edges E, where each edge $e \in E$ contains r distinct vertices from V. So a graph is a 2-uniform hypergraph. Szemerédi's famous theorem [35] extends Roth's theorem by showing that for every fixed k, if $S \subseteq [n]$ contains no k-term arithmetic progression then $|S| = o(n)$. Motivated by the fact that a removal lemma for graphs can be used to prove Roth's theorem, Frankl and Rödl [13] showed that a removal lemma for r-uniform hypergraphs could be used to prove Szemerédi's theorem on $(r+1)$-term arithmetic progressions. They further developed a regularity lemma, as well as a corresponding removal lemma, for 3-uniform hypergraphs thus obtaining a new proof of Szemerédi's theorem for 4-term arithmetic progressions. In recent years there have been many exciting results in this area, in particular the results of Gowers [16] and of Nagle, Rödl, Schacht and Skokan [25, 26], who independently obtained regularity lemmas and removal lemmas for r-uniform hypergraphs, thus providing alternative combinatorial proofs of Szemerédi's Theorem [35] and some of it generalizations, notably those of Furstenberg and Katznelson [14]. Tao [37] and Ishigami [19] later obtained another proof of the hypergraph removal lemma and of its many corollaries mentioned above. For more details see [17, 21].

In this paper we will use the above mentioned hypergraph removal lemma in order to resolve a conjecture of Green [18] regarding the removal properties of sets of linear equations. Let $Mx = b$ be a set of linear equations, and let us say that a set of integers S is (M, b)-free if it contains no solution to $Mx = b$, that is, if there is no vector x, whose entries all belong to S, which satisfies $Mx = b$. Just as the removal lemma for graphs states that a graph that has few copies of H is close to being H-free, a removal lemma for sets of linear equations $Mx = b$ should say that a subset of the integers $[n]$ that contains few solutions to $Mx = b$ is close to being (M, b)-free. Let us start by defining this notion precisely.

Definition 1.1 (Removal Property) *Let M be an $\ell \times p$ matrix of integers and let $b \in \mathbb{N}^{\ell}$. The set of linear equations $Mx = b$ has the* removal property *if for every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, M, b) > 0$ with the following property: if $S \subseteq [n]$ is such that there are at most $\epsilon n^{p-\ell}$ vectors $x \in S^p$ satisfying $Mx = b$, then one can remove from S at most δn elements to obtain an (M, b)-free set.*

We note that in the above definition, as well as throughout the paper, we assume that the $\ell \times p$ matrix M of a set of linear equations has rank ℓ.

Green [18] has initiated the study of the removal properties of sets of linear equations. His main result was the following:

Theorem 1 (Green [18]) *Any single homogenous linear equation has the removal property.*

The main result of Green actually holds over any abelian group. To prove this result, Green developed a regularity lemma for abelian groups, which is somewhat analogous to Szemerédi's regularity lemma for graphs [36]. Although the application of the group regularity lemma for proving Theorem 1 was similar to the derivation of the graph removal lemma from the graph regularity lemma, the proof of the group regularity lemma was far from trivial. One of the main conjectures raised in [18] is that a natural generalization of Theorem 1 should also hold (Conjecture 9.4 in [18]).

Conjecture 1 (Green [18]) *Any system of homogenous linear equations $Mx = 0$ has the removal property.*

We note that besides being a natural generalization of Theorem 1, Conjecture 1 was also raised in [18] with relation to a conjecture of Bergelson, Host, Kra and Ruzsa [7] regarding the number of k-term arithmetic progressions with a common difference in subsets of $[n]$. See Section 5 for more details.

1.2 Recent progress on Green's Conjecture and our main result

Very recently, Král', Serra and Vena [23] gave a surprisingly simple proof of Theorem 1, which completely avoided the use of Green's regularity lemma for groups. In fact, their proof is an elegant and simple application of the removal lemma for directed graphs [2], which is a simple variant of the graph removal lemma that we have previously discussed. The proof given in [23] actually extends Theorem 1 to any single non-homogenous linear equation over arbitrary groups. Král', Serra and Vena [23] also show that Conjecture 1 holds when M is a 0/1 matrix, which satisfies certain conditions. But these conditions are not satisfied even by all 0/1 matrices.

Let us now mention some results that have all been obtained independently of ours. Bhattacharyya, Chen, Sudan and Xie [9] showed that Conjecture 1 holds when the system of equations can be realized as a graphical matroid, see [9] for the exact details. The proof of the main result of [9] is mainly analytic, and applies Green's regularity lemma for groups [18] mentioned above. Bhattacharyya et al. stated that extending their result to the full generality of Conjecture 1 "seems to pose significant technical hurdles". They further posed as a challenging open problem to show that the system of homogenous linear equations $x_1 + x_2 + x_4 = 0$, $x_2 + x_3 + x_5 = 0$, $x_3 + x_1 + x_6 = 0$, $x_1 + x_2 + x_3 + x_7 = 0$ has the removal property. In another recent result, Candela [11] showed that Conjecture 1 holds for every system of ℓ homogenous linear equations $Mx = 0$ in which every ℓ columns of M are linearly independent. See more details in Subsection 2.1. A similar result to Candela's [11], was also obtained by Král', Serra and Vena [24].

In this paper we confirm Green's conjecture for every homogenous set of linear equations. In fact, we prove the following more general result.

Theorem 2 (Main Result) *Any set of linear equations (even non homogenous) $Mx = b$ has the removal property.*

1.3 Applications to testing properties of boolean functions

Besides being a natural problem from the perspective of additive number theory, it turns out that Theorem 2 has some applications in Theoretical Computer Science, in the area of *Property Testing.* Property testers are fast randomized algorithms that can distinguish between objects satisfying a certain property $\mathcal{P}$ and objects that are "far" from satisfying it. In an attempt to prove a general sufficient condition that would guarantee that certain properties of boolean functions have efficient testing algorithms, Bhattacharyya, Chen, Sudan and Xie [9] conjectured that certain properties of boolean functions (that are related to the notion of being (M, b)-free) can be efficiently tested. As we show in Section 4, our main result gives a positive answer to their open problem.

1.4 Organization

The rest of the paper is organized as follows. In the next section we give an overview of the proof of Theorem 2. As we show in that section, Theorem 2 also holds over any finite field, that is when $S \subseteq \mathbb{F}_n$, where $\mathbb{F}_n$ is the field of size n. In fact it is easy to modify the proof so that it works over any field, but we will not do so here. The proof of Theorem 2 has two main steps: the first one, described in Lemma 2.3, applies the main idea from [23] in order to show that if a set of linear equations can be "represented" by a hypergraph then Theorem 2 would follow from a variant of the hypergraph removal lemma. So the second, and most challenging, step of the proof is showing that every set of linear equations can be represented as a hypergraph. The proof of this result, stated in Lemma 2.4, appears in Section 3. In Section 4 we describe the applications of our main result to devising efficient testing algorithms for properties of boolean functions. Finally, in Section 5 we give some concluding remarks and discuss some open problems.

More recent results: After our paper appeared on the arXiv we have learned that independently of our work here, Král', Serra and Vena managed to improve upon their results in [23, 24] and obtain a proof of Conjecture 1.

2 Proof Overview

It will be more convenient to deduce Theorem 2 from an analogous result over the finite field $\mathbb{F}_n$ of size n (for n a prime power). In fact, somewhat surprisingly, we will actually need to prove a stronger claim than the one asserted in Theorem 2. This more general variant, stated in Theorem 3, allows each of the variables x_i to have its own subset $S_i \subseteq \mathbb{F}_n$. We note that a proof of this variant of Theorem 2 for the case of a single equation was already proved in [18] and [23], but in those papers it was not necessary to go through this more general result. As we will explain later (see Claims 3.1 and 3.3), the fact that we are considering a more general problem will allow us to overcome some degeneracies in the system of equations by allowing us to remove certain equations. This manipulation can be performed when one considers the generalized removal property (defined below) but there is no natural way of performing these manipulations when considering the standard removal property. Therefore, proving this extended result is essential for our proof strategy.

In what follows and throughout the paper, whenever x is a vector, x_i will denote its i^{th} entry. Similarly, if $x_1, \ldots, x_p$ are elements in a field, then x will be the vector whose entries are $x_1, \ldots, x_p$. We say that a collection of p subsets $S_1, \ldots, S_p \subseteq \mathbb{F}_n$ is (M, b)-free if there are no $x_1 \in S_1, \ldots, x_p \in S_p$ which satisfy $Mx = b$.

Definition 2.1 (Generalized Removal Property over Finite Fields) *Let $\mathbb{F}_n$ be the field of size n, let M be an $\ell \times p$ matrix over $\mathbb{F}_n$ and let $b \in \mathbb{F}_n^{\ell}$. The system $Mx = b$ is said to have the* generalized removal property *if for every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, p) > 0$ such that if $S_1, \ldots, S_p \subseteq \mathbb{F}_n$ contain less than $\epsilon n^{p-\ell}$ solutions to $Mx = b$ with each $x_i \in S_i$, then one can remove from each S_i at most δn elements to obtain sets $S'_1, \ldots, S'_p$ which are (M, b)-free.*

By taking all sets S_i to be the same set S we, of course, get the standard notion of the removal property from Definition 1.1 so we may indeed work with this generalized definition. We will deduce Theorem 2 from the following theorem.

Theorem 3 *Every set of linear equations $Mx = b$ over a finite field has the generalized removal property.*

In order to prove Theorem 3 we will need to apply the hypergraph removal lemma. We will actually need a variant of the hypergraph removal lemma which works for colored hypergraphs. Let us first recall some basic definitions. An r-uniform hypergraph is simple if it has no parallel edges, that is, if different edges contain different subsets of vertices of size r. We say that a set of vertices U in an r-uniform hypergraph $H = (V_H, E_H)$ spans a copy of an r-uniform hypergraph $K = (V_K, E_K)$ if there is an injective mapping ϕ from V_K to U such that if $v_1, \ldots, v_r$ form an edge in K then $\phi(v_1), \ldots, \phi(v_r)$ form an edge in $U \subseteq V_H$. We say that a hypergraph is c-colored if its edges are colored by $\{1, \ldots, c\}$. If K and H are c-colored, then U is said to span a colored copy of K if the above mapping ϕ sends edges of K of color i to edges of H (in U) of the same color i. We stress that the coloring of the edges does not have to satisfy any constraints that are usually associated with edge colorings. Finally, the number of colored copies of K in H is the number of subsets $U \subseteq V_H$ of size $|V_K|$ which span a colored copy of K.

The following variant of the hypergraph removal lemma was proved in [19] Theorem 4.1, and is also a special case of Theorem 1.2 in [5].[1]

Theorem 4 (Austin and Tao [5], Ishigami [19]) *Let K be a fixed r-uniform c-colored hypergraph on k vertices. For every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, k) > 0$ such that if H is an r-uniform c-colored simple hypergraph with less than ϵn^k colored copies of K, then one can remove from H at most δn^r edges and obtain a hypergraph that contains no colored copy of K.*

In order to use Theorem 4 for the proof of Theorem 3, we will need to represent the solutions of $Mx = b$ as colored copies of a certain "small" hypergraph K in a certain "large" hypergraph H. The following notion of hypergraph representability specifies the requirements from such a representation that suffice for allowing us to deduce Theorem 3 from Theorem 4.

Definition 2.2 (Hypergraph Representation) *Let $\mathbb{F}_n$ be the field of size n, let M be an $\ell \times p$ matrix over $\mathbb{F}_n$. The system of linear equations $Mx = b$ is said to be* hypergraph representable *if there is an integer $r = r(M, b) \leq p^2$ and an r-uniform p-colored hypergraph K with $k = r - 1 + p - \ell$ vertices and p edges, such that for any $S_1, \ldots, S_p \subseteq \mathbb{F}_n$ there is an r-uniform hypergraph H on kn vertices which satisfies the following:*

1. *H is simple and each edge with color i is labeled by one of the elements of S_i.*

2. *If $x_1 \in S_1, \ldots, x_p \in S_p$ satisfy $Mx = b$ then H contains n^{r-1} p-colored copies of K. In each of these copies, the edge with color i has the label x_i. These colored copies of K should also be edge disjoint.*

3. *If $S_1, \ldots, S_p$ contain at most T solutions to $Mx = b$ with $x_i \in S_i$ then H contains at most Tn^{r-1} colored copies of K.*

We note that the notion of hypergraph representation we defined above is a variant of other representations that have been previously used in order to obtain some previous result. Ruzsa and Szemerédi [30], in their proof of Roth's Theorem [27], where the first to represent the solution of a linear equation using a graph. Frankl and Rödl [13] used an extension of the Ruzsa and Szemerédi construction in their hypergraph approach to the proof of Szemerédi's Theorem. Finally, the recent proof of Král et al. [23] that a single linear equation has the removal property used another variant of this representation. Although our proof here is self contained a reader who is unfamiliar with this approach may benefit from consulting the short proof in [23].

The following lemma shows that a hypergraph representation can allow us to prove Theorem 3 using the hypergraph removal lemma.

[1] As noted to us by Terry Tao, this variant of the hypergraph removal lemma can probably be extracted from the previous proofs of the hypergraph removal lemma [16, 25, 26, 37], just like the colored removal lemma for graphs can be extracted from the proof of the graph removal lemma, see [22].

Lemma 2.3 *If $Mx = b$ has a hypergraph representation then it has the generalized removal property.*

Proof: Suppose $Mx = b$ is a system of ℓ linear equations in p unknowns. Let $S_1, \ldots, S_p$ be p subsets of $\mathbb{F}_n$ and let $H = (V, E)$ be the hypergraph guaranteed by Definition 2.2. We claim that we can take $\epsilon(\delta, p)$ in Definition 2.1 to be the value $\epsilon = \epsilon(\delta/pk^r, k)$ from Theorem 4. Indeed, if $S_1, \ldots, S_p$ contain only $\epsilon n^{p-\ell}$ solutions to $Mx = b$ then by item 3 of Definition 2.2 we get that H contains at most $\epsilon n^{p-\ell} \cdot n^{r-1} = \epsilon n^k$ colored copies of K. As H is simple[2], we can apply the removal lemma for colored hypergraphs (Theorem 4) to conclude that one can remove a set $\widetilde{E}$ of at most $\frac{\delta}{pk^r}(kn)^r = \frac{\delta}{p}n^r$ edges from H and thus destroy all the colored copies of K in H (recall that H has kn vertices). Note that since $r, k \leq 2p^2$, we have that $\epsilon = \epsilon(\delta/pk^r, k)$ is bounded by a function of δ and p only, as required by Definition 2.1.

To show that we can turn $S_1, \ldots, S_p$ into a collection of (M, b)-free sets by removing only δn elements from each S_i, let us remove an element s from S_i if $\widetilde{E}$ contains at least n^{r-1}/p edges that are colored with i and labeled with s. As each edge has one label (because H has no parallel edges), and $|\widetilde{E}| \leq \frac{\delta}{p}n^r$ this means that we remove only δn elements from each S_i. To see that we thus turn $S_1, \ldots, S_p$ into (M, b)-free sets, suppose that the new sets $S'_1, \ldots, S'_p$ still contain a solution $s_1 \in S_1, \ldots, s_p \in S_p$ to $Mx = b$. By item 2 of Definition 2.2, this solution defines n^{r-1} edge disjoint p-colored copies of K in H, with the property that in every colored copy, the edge with color i is labeled with the same element $s_i \in S_i$. As $\widetilde{E}$ must contain at least one edge from each of these colored copies (as it should destroy all such copies), there must be some $1 \leq i \leq p$ for which $\widetilde{E}$ contains at least n^{r-1}/p edges that are colored i and labeled with s_i. But this contradicts the fact that s_i should have been removed from S_i. ■

We note that the above lemma generalizes a similar lemma for the case of representing a single equation using a graph, which was implicit in [23]. In fact, as we have mentioned earlier, [23] also show that a set of homogenous linear equations $Mx = 0$, with M being a 0/1 matrix, that satisfies certain conditions also has the removal lemma. One of these conditions essentially says that the system of equations is *graph* representable. However, there are even some 0/1 matrices for which $Mx = 0$ is not graph representable (in the sense of [23]). Lemma 2.4 below shows that any set of linear equations has a *hypergraph* representation. This lemma is proved in the next section and it is the most challenging part of this paper.

Lemma 2.4 *Every set of linear equations $Mx = b$ over a finite field is hypergraph representable.*

The above two lemmas give the following.

[2] We note that it is important to require H to be simple since in general one cannot apply Theorem 4 to graphs (or hypergraphs) with parallel edges, see [33].

Proof of Theorem 3: Immediate from Lemmas 2.3 and 2.4.

As we have mentioned before, Theorem 2 is now an easy application of Theorem 3.

Proof of Theorem 2: Given a set of linear equations $Mx = b$ in p unknowns, let c be the maximum absolute value of the entries of M and b. Given an integer n let $q = q(n)$ be the smallest prime larger than cp^2n. It is well known that $q \leq 2cp^2n$ (in fact, much better bounds are known). It is clear that for a vector $x \in [n]^p$ we have $Mx = b$ over $\mathbb{R}$ if and only if $Mx = b$ over $\mathbb{F}_q$. So if $Mx = b$ has $o(n^{p-\ell})$ solutions with $x_i \in S_i$ over $\mathbb{R}$, it also has $o(q^{p-\ell})$ solutions with $x_i \in S_i \subseteq \mathbb{F}_q$ over $\mathbb{F}_q$. By Theorem 3 we can remove $o(q)$ elements from each S_i and obtain sets S'_i that are (M, b)-free. But as $q = O(n)$ we infer that the removal of the same $o(q) = o(n)$ elements also guarantees that the sets are (M, b)-free over $\mathbb{R}$. ■

2.1 Overview of the Proof of Lemma 2.4

Let us start by noting that Lemma 2.4 for the case of a single equation was (implicitly) proven in [23], where they show that one can take $r = 2$, in other words, they represent a single equation as a graph K, in a graph H. Actually, the graph K in the proof of [23] is a cycle of length p. The proof in [23] is very short and elegant, and we recommend reading it to better understand the intuition behind our proof. Another related result is the proof of Szemerédi's theorem [35] using the hypergraph removal lemma [13], which can be interpreted as (essentially) showing that the set of $p - 2$ linear equations which define a p-term arithmetic progression[3] are hypergraph representable with K being the complete $(p-1)$-uniform hypergraph of size p. "Interpolating" these two special cases of Lemma 2.4 suggests that a hypergraph representation of a set of ℓ linear equations in p unknowns should involve an $(\ell + 1)$-uniform hypergraph K of size p. And indeed, we initially found a (relatively) simple way to achieve this for $p - 2$ equations in p unknowns, thus extending the representability of the arithmetic progression set of linear equations.

However, somewhat surprisingly, when $1 < \ell < p - 2$ the situation becomes much more complicated and we did not manage to find a simple representation along the lines of the above two cases. The problem with trying to extend the previous approaches to larger sets of equations is that obtaining all the requirements of Definition 2.2 turns out to be very complicated when M has a set of ℓ columns that are not linearly independent. Let us mention again that Candela [11] has recently considered linear equations $Mx = 0$ in which every ℓ columns are linearly independent, and showed that Conjecture 1 holds in these cases.

The way we overcome the above complications is by using a representation involving hypergraphs of a much larger degree of uniformity (that is, larger edges), which is roughly the number of non-zero entries of M after we perform certain manipulations on it. We note that specializing our proof to

[3] These linear equations are $x_1 + x_3 = 2x_2$, $x_2 + x_4 = 2x_3, \ldots, x_{p-2} + x_p = 2x_{p-1}$.

either the case $\ell = 1$ or to the case $\ell = p - 2$ does *not* give proofs that are identical to the ones (implicit) in [13] or [23]. For example, our proof for the case of a single equation in p unknowns uses a $(p-1)$-uniform hypergraph, rather than a graph as in [23].

So let us give a brief overview of the proof. Given a set of ℓ linear equations in p unknowns, we need to find a "small" hypergraph K with p edges, whose copies, within another "large" hypergraph H, will represent the solutions to $Mx = b$. Each edge of H, and therefore also K, will have a color $1 \leq i \leq p$ and a label $s \in S_i$. The system $Mx = b$ has p unknowns and K has p edges and it may certainly be the case that all the entries of M are non-zero. It is apparent that using all the edges of K to "deduce" a linear equation of $Mx = b$ is not a good idea because in that way we will only be able to extract one equation from a copy of K and we need to extract ℓ such equations. Therefore, we will first "diagonalize" an $\ell \times \ell$ sub-matrix of M to get an equivalent set of equations (which we still denote by $Mx = b$) which has the property that $p - \ell$ of its unknowns $x_1, \ldots, x_{p-\ell}$ (can) appear in all the equations and the rest of the ℓ unknowns $x_{p-\ell+1}, \ldots, x_p$ each appear in precisely one equation. This suggests the idea of extracting equation i from (some of) the edges corresponding to $x_1, \ldots, x_{p-\ell}$ and one of the edges corresponding to $x_{p-\ell+1}, \ldots, x_p$. The hypergraph K first contains $p - \ell$ edges that do not depend on the structure of M. The other ℓ edges do depend on the structure of M and use the previous $p - \ell$ edges in order to "construct" the equations of $Mx = b$. The way to think about this is that for any copy of K in H the first $p - \ell$ edges will have a special vertex that will hold a value from S_i (this will be the vertex in one of the sets $U_1, \ldots, U_{p-\ell}$ defined in Section 3). The other ℓ edges will include some of these special vertices, depending on the equation we are trying to build. The way we will deduce an equation from a copy of K in H is that we will argue that the fact that two edges have a common vertex means that a certain equation holds. See Claim 3.4.

But there is another complication here because the linear equation we obtain in the above process will contain many other variables not from the sets S_i, which will need to vanish from such an equation, in order to allow us to extract the linear equations we are really interested in. The reason for these "extra" variables is that by item (2) of Definition 2.2, H needs to contain n^{r-1} edge disjoint copies of K for every solution of $Mx = b$. Hence, an edge of H will actually be parameterized by several other elements from $\mathbb{F}_n$ (these are the elements $x_1, \ldots, x_{r-1}$ that are used after Claim 3.2). So we will need to make sure that these extra variables vanish in the linear equation which we extract from a copy of K. To make sure this happens we will need to carefully choose the vertices of each edge within H.

A final complication arises from the fact that while we want H to contain relatively few copies of K (item (3) of Definition 2.2), we also want it to contain many edge disjoint copies of K for every solution of $Mx = b$ (item (2) of Definition 2.2). To this end we will think of each vertex of H as a linear equation and we will want the linear equations corresponding to the vertices of an edge to be linearly independent. The reason why it is hard to prove Lemma 2.4 using an $(\ell + 1)$-

uniform hypergraph (as the results of [23] and [13] may suggest) is that it seems very hard to obtain all the above requirements simultaneously. The fact that we are considering hypergraphs with a larger degree of uniformity will allow us (in some sense) to break the dependencies between these requirements.

3 Proof of Lemma 2.4

Let $Mx = b$ be the set of linear equation, where M is an $\ell \times p$ matrix over $\mathbb{F}_n$ and $b \in \mathbb{F}_n^{\ell}$. We will first perform a series of operations on M and b which will help us in proving Lemma 2.4. For convenience, we will continue to refer to the transformed matrix and vector as M and b. Suppose, without loss of generality, that the last ℓ columns of M are linearly independent. We can thus transform M (and accordingly also b) into an equivalent set of equations in which the last ℓ columns form an identity matrix. For a row M_i of M let m_i be the largest index $1 \leq j \leq p - \ell$ for which $M_{i,j}$ is non-zero. Let W_i denote the set of indices $1 \leq j \leq m_i - 1$ for which $M_{i,j}$ is non-zero. Therefore, M_i has $|W_i| + 2$ non-zero entries. We will need the following claim, in which we make use of the fact that we are actually proving that every set of equations has the generalized removal property and not just the removal property.

Claim 3.1 *Suppose that every set of $\ell - 1$ equations in $p - 1$ unknowns over $\mathbb{F}_n$ has the generalized removal property. Suppose that the matrix M defined above has a row with less than 3 non-zero entries. Then $Mx = b$ has the generalized removal property as well.*

Proof: Suppose that (say) the first row of M has at most 2 non-zero entries. If this row has two non-zero elements then we can assume without loss of generality that it is of the form $x_1 = q - a \cdot x_j$ where $p - \ell + 1 \leq j \leq p$. But then we can get an equivalent set of linear equations $M'x = b'$ by removing the first row from M, removing the column in which x_j appears (because x_j does not appear in other rows), removing the first entry of b and updating S_1 to be $S_1' = S_1 \cap \{q - a \cdot s \ : \ s \in S_j\}$. We thus get an instance $M'x = b'$ with $\ell - 1$ equations and $p - 1$ unknowns, hence we can use the assumption of the claim because: (i) The number of solutions of $Mx = b$ with $x_i \in S_i$ is precisely the number of solutions of $M'x = b'$ with $x_1 \in S_1', x_2 \in S_2, \ldots, x_{j-1} \in S_{j-1}, x_{j+1} \in S_{j+1}, \ldots, x_p \in S_p$ (ii) if we can remove δn elements from each of the sets of the new instance and thus obtain sets with no solution of $M'x = b'$ then the removal of the same elements from the original sets S_i would also give sets with no solution of $Mx = b$.

If the first row of M has just one non-zero entry, then this equation is of the form $x_j = q$ for some $p - \ell + 1 \leq j \leq p$ and $q \in \mathbb{F}_n$. If $q \notin S_j$ then the sets contain no solution to $Mx = b$ and there is nothing to prove. If $q \in S_j$ then the number of solutions to $Mx = b$ is the number of solutions of the set of equations $M'x = b'$ where M' is obtained by removing the row and column to which x_j

belongs and by removing the first entry of b. As in the previous case we can now use the assumption of the claim. ■

Claim 3.1 implies that we can assume without loss of generality that none of the sets $W_1, \ldots, W_\ell$ is empty, because if one of them is empty then the corresponding row of M contains less than 3 non-zero entries. In that case we can iteratively remove equations from M until we either: (i) get a set of linear equations in which none of the rows has less than 3 non-zero entries, in which case we can use the fact that the result holds for such sets of equations as we will next show, or (ii) we get a single equation with only 2 unknowns with a non-zero coefficient[4]. It is now easy to see that such an equation has the removal property. Indeed, suppose the equation has p unknowns and only x_1 and x_2 have a non-zero coefficient. So the equation is $a_1 \cdot x_1 + a_2 \cdot x_2 + \sum_{i=3}^{p} 0 \cdot x_i = b$. In this case the number of solutions to the equation from sets $S_1, \ldots, S_p$ is the number of solutions to the equation $a_1 x_1 + a_2 x_2 = b$ with $x_1 \in S_1, x_2 \in S_2$ multiplied by $\prod_{i=3}^{p} |S_i|$. Therefore, if $S_1, \ldots, S_p$ contain $o(n^{p-1})$ solutions, then either (i) one of the sets $S_3, \ldots, S_p$ is of size $o(n)$, so we can remove all the elements from this set, or (ii) S_1, S_2 contain $o(n)$ solutions to $a_1 \cdot x_1 + a_2 \cdot x_2 = b$, but in this case, for every solution (s_1, s_2) we can remove s_1 from S_1. In either case the new sets $S'_1, \ldots, S'_p$ contain no solution of the equation, as needed.

We now return to the proof of Lemma 2.4, with the assumption that none of the sets W_i is empty. Let us multiply each of the rows of M by M^{-1}_{i,m_i} so that for every $1 \leq i \leq \ell$ we have $M_{i,m_i} = 1$. For every $1 \leq i \leq \ell$ let $d_i \in \{p-\ell+1, \ldots, p\}$ denote the index of the unique non-zero entry of M_i within the last ℓ columns of M. Using the notation which we have introduced thus far, the system of linear equations $Mx = b$ can be written as the set of ℓ equations $L_1, \ldots, L_\ell$, where L_i is the equation

$$x_{m_i} + M_{i,d_i} \cdot x_{d_i} + \sum_{j \in W_i} M_{i,j} \cdot x_j = b_i \ . \tag{1}$$

Let us set

$$r = 1 + \sum_{1 \leq i \leq \ell} |W_i| \ .$$

Observe that as required by Definition 2.2 we indeed have $r \leq p^2$.

We now define an r-uniform p-colored hypergraph K, which will help us in proving that $Mx = b$ is hypergraph representable as in Definition 2.2. The hypergraph K has $k = r - 1 + p - \ell$ vertices (as required by Definition 2.2) which we denote by $v_1, \ldots, v_{r-1}, u_1, \ldots, u_{p-\ell}$. As for K's edges, it first contains $p - \ell$ edges denoted $e_1, \ldots, e_{p-\ell}$, where e_i contains the vertices $v_1, \ldots, v_{r-1}, u_i$. Note that these edges do not depend on the system $Mx = b$. As we will see later, these edges will help us to "build" the actual representation of the linear equations of $Mx = b$. So in addition to the above $p - \ell$ edges, K also contains ℓ edges $f_{p-\ell+1}, \ldots, f_p$, where edge f_{d_i} will[5] represent (in some sense)

[4] Note that this process can result in having unknowns with a zero coefficient in all the remaining equations.

[5] Note that we are using the fact that $d_1, \ldots, d_\ell$ are distinct numbers in $\{p - \ell + 1, \ldots, p\}$.

equation L_i, defined in (1). To define these ℓ edges it will be convenient to partition the set $[r-1]$ into ℓ subsets $I_1, \ldots, I_\ell$ such that I_1 contains the numbers $1, \ldots, |W_1|$, and I_2 contains the numbers $|W_1|+1, \ldots, |W_1|+|W_2|$ and so on. With this partition we define for every $1 \leq i \leq \ell$ edge f_{d_i} to contain the vertices $\{v_j \ : \ j \in [r-1] \setminus I_i\}$, the vertices $\{u_j \ : \ j \in W_i\}$ as well as vertex u_{m_i}. Note that as $|I_i| = |W_i|$ the hypergraph K is indeed r-uniform. As for the coloring of the edges of K, for every $1 \leq i \leq p-\ell$ edge e_i is colored i and for every $p-\ell+1 \leq d_i \leq p$ edge f_{d_i} is colored d_i.

We now turn to define certain $p-\ell$ vectors $a^1, \ldots, a^{p-\ell} \in \mathbb{F}_n^{r-1}$ which will be used later in the construction of the hypergraph H. We think of $a^1, \ldots, a^{p-\ell}$ as the $p-\ell$ rows of a $(p-\ell) \times (r-1)$ matrix A. Furthermore, for every $1 \leq i \leq \ell$ let A_i be the sub-matrix of A which contains the columns whose indices belong to I_i (which was defined above). We now take the (square) sub-matrix of A_i which contains the rows whose indices belong to W_i to be the identity matrix (over $\mathbb{F}_n$). More precisely, if the elements of W_i are $j_1 < j_2 < \ldots < j_{|W_i|}$ then $A'_{j_g,g} = 1$ for every $1 \leq g \leq |W_i|$, and 0 otherwise[6]. For future reference, let's denote by A'_i this square sub-matrix of A_i. We finally set row m_i of A_i to be the vector whose g^{th} entry is $-M_{i,j_g}$, where as above j_g is the g^{th} element of W_i. If A_i has any other rows besides the ones defined above, we set them to 0. As each column of A belongs to one of the matrices A_i we have thus defined A and therefore also the vectors $a^1, \ldots, a^{p-\ell}$.

Let us make two simple observations regarding the above defined vectors which we will use later. First, let $1 \leq i \leq \ell$ and $t \in I_i$ and suppose t is the g^{th} element of I_i. Then[7]

$$\sum_{j \in W_i} a^j_t \cdot M_{i,j} = (A_i)_{j_g,g} \cdot M_{i,j_g} = M_{i,j_g} = -(A_i)_{m_i,g} = -a^{m_i}_t \ , \tag{2}$$

where the first equality is due to the fact that the only non-zero entries within column g of A_i and the rows from W_i appears in row j_g. The second equality uses the fact that this entry is in fact 1. The third equality uses the definition of row m_i of A_i.

The second observation we will need is the following.

Claim 3.2 *For $1 \leq i \leq \ell$, let B_i be the following $r-1 \times r-1$ matrix: for every $j \in [r-1] \setminus I_i$ we have $(B_i)_{j,j} = 1$ and $(B_i)_{j,t} = 0$ for $t \neq j$. The other $|I_i|$ rows of B_i are the $|W_i|$ $(= |I_i|)$ vectors $\{a^t : t \in W_i\}$. Then, for every $1 \leq i \leq \ell$ the matrix B_i is non-singular.*

Proof: To show that B_i is non-singular it is clearly enough to show that its $|I_i| \times |I_i|$ minor B'_i, which is determined by I_i, is non-singular. But observe that this fact follows from the way we have defined the vectors $a^1, \ldots, a^{p-\ell}$ above because B'_i is just A'_i, which is in fact the identity matrix. ■

We are now ready to define, for every set of subsets $S_1, \ldots, S_p \subseteq \mathbb{F}_n$, the hypergraph H which will establish that $Mx = b$ is hypergraph representable. The vertex set of H consists of k $(= r-1+p-\ell)$

[6] Note that the second index of $A'_{j_g,g}$ refers to the column number within A_i, not A.

[7] Note that t is an index of a column of A, while g is an index of a column of A_i.

disjoint sets $V_1, \dots, V_{r-1}, U_1, \dots, U_{p-\ell}$, where each of these sets contains n vertices and we think of the elements of each of these sets as the elements of $\mathbb{F}_n$. As for the edges of H, we first put for $1 \le i \le p - \ell$ and every choice of $r-1$ vertices $x_1 \in V_1, \dots, x_{r-1} \in V_{r-1}$ and element $s \in S_i$, an edge with color i and label s, which contains the vertices $x_1, \dots, x_{r-1}$ as well as vertex $y \in U_i$, where

$$y = s + \sum_{j=1}^{r-1} a^i_j x_j \ , \tag{3}$$

and the values a^i_j were defined above. These edges will later play the role of the edges $e_1, \dots, e_{p-\ell}$ of K defined above. Note that these edges are defined irrespectively of the set of equations $Mx = b$.

We now define the edges of H which will "simulate" the linear equations of $Mx = b$. For every $1 \le i \le \ell$, and for every choice of an element $s \in S_{d_i}$, for every choice of $r - 1 - |I_i|$ vertices $\{x_t \in V_t \ : \ t \in [r-1] \setminus I_i\}$ and for every choice of $|W_i|$ $(= |I_i|)$ vertices $\{y_j \in U_j \ : \ j \in W_i\}$ we have an edge with color d_i and label s, which contains the vertices $\{x_t : t \in [r-1] \setminus I_i\}$ and $\{y_j \ : \ j \in W_i\}$ as well as vertex $y \in U_{m_i}$, where

$$y = b_i - M_{i,d_i} \cdot s - \sum_{j \in W_i} M_{i,j} \cdot y_j + \sum_{t \in [r-1] \setminus I_i} x_t \cdot (a^{m_i}_t + \sum_{j \in W_i} a^j_t \cdot M_{i,j}) \ . \tag{4}$$

Let us first note that as required by Lemma 2.4, each edge of H has a color i and is labeled by an element $s \in S_i$. In fact, for each $1 \le i \le p$ and for each $s \in S_i$, the hypergraph H has n^{r-1} edges that are colored i and labeled with s. We now turn to establish the first property required by Definition 2.2.

Claim 3.3 *H is a simple hypergraph, that is, it contains no parallel edges.*

Proof: Observe that edges of H with different colors have a single vertex from a different subset of r of the sets $V_1, \dots, V_{r-1}, U_1, \dots, U_{p-\ell}$. Indeed, edges with color $1 \le i \le p - \ell$ contain a vertex from each of the sets $V_1, \dots, V_{r-1}$ and another vertex from U_i, while an edge with color $p - \ell + 1 \le d_i \le p$ contains vertices from the sets $\{V_t \ : \ t \in [r-1] \setminus I_i\}$ as well as vertices from some of the sets $U_1, \dots, U_{p-\ell}$. Note that the sets $I_1, \dots, I_\ell$ are disjoint and non-empty, as none of the sets W_i is empty, a fact which (as noted previously) follows from Claim 3.1. Observe that if W_i was empty, then edges with color d_i would have had parallel edges with color m_i.

As for edges with the same color $1 \le i \le p - \ell$, recall that they are defined in terms of a different combination of $x_1, \dots, x_{r-1} \in \mathbb{F}_n$ and $s \in S_i$. So if one edge is defined in terms of $x_1, \dots, x_{r-1} \in \mathbb{F}_n$ and $s \in S_i$ and another using $x'_1, \dots, x'_{r-1} \in \mathbb{F}_n$ and $s' \in S_i$ then consider the following two cases; (i) $x_j \ne x'_j$ for some $1 \le j \le r-1$: in this case the edges have a different vertex in V_j; (ii) $x_j = x'_j$ for all $1 \le j \le r-1$: In this case $s \ne s'$. Therefore the edges have a different vertex in U_i by the way we chose the vertex in this set in (3).

The case of edges with the same color $p - \ell + 1 \leq d_i \leq p$ is similar. Recall that such edges are defined in terms of a different combination of $\{x_t \ : \ t \in [r-1] \setminus I_i\}$, $\{y_j \ : \ j \in W_i\}$ and $s \in S_{d_i}$. So if one edge is defined in terms of $\{x_t \ : \ t \in [r-1] \setminus I_i\}$, $\{y_j \ : \ j \in W_i\}$ and $s \in S_{d_i}$ and another using $\{x'_t \ : \ t \in [r-1] \setminus I_i\}$, $\{y'_j \ : \ j \in W_i\}$ and $s' \in S_{d_i}$ then consider the following three cases; (i) $x_t \neq x'_t$ for some $t \in [r-1] \setminus I_i$: in this case the edges have a different vertex in V_t; (ii) $y_j \neq y'_j$ for some $j \in W_i$: in this case the edges have a different vertex in U_j (iii) $x_t = x'_t$ for all $t \in [r-1] \setminus I_i$, and $y_j = y'_j$ for all $j \in W_i$: In this case $s \neq s'$ and therefore the edges have a different vertex in U_{m_i} by the way we chose the vertex in this set in (4) and from the fact that we chose d_i in away that $M_{i,d_i} \neq 0$ (recall the paragraph preceding equation (1)). ■

We now turn to establish the second and third properties required by Definition 2.2. Fix arbitrary elements $s_1 \in S_1, \ldots, s_{p-\ell} \in S_{p-\ell}$. For every choice of $r-1$ (not necessarily distinct) elements $x_1, \ldots x_{r-1} \in \mathbb{F}_n$, let K_x be the set of vertices $x_1 \in V_1, \ldots, x_{r-1} \in V_{r-1}, y_1 \in U_1, \ldots, y_{p-\ell} \in U_{p-\ell}$, where for every $1 \leq j \leq p - \ell$

$$y_j = s_j + \sum_{t=1}^{r-1} a_t^j \cdot x_t \ . \tag{5}$$

We will need the following important claim regarding the vertices of K_x. Getting back to the overview of the proof given in Subsection 2.1, this is where we extract one of the linear equations L_i (defined above) from a certain combination of edges of a copy of K. We also note that the linear equation we "initially" obtain (see (6)) includes also the elements x_i, but the way we have constructed H guarantees that the x_i's vanish and we eventually get a linear equation involving only elements from the sets S_i. We will then use this claim to show that H contains many edge disjoint copies of K when $s_1, \ldots, s_{p-\ell}$ determine a solution to $Mx = b$, and in the other direction, that H cannot contain too many copies of K. For what follows we remind the reader that for $1 \leq i \leq \ell$ we have $p - \ell + 1 \leq d_i \leq p$ and that for $i < i'$ we have $d_i \neq d_{i'}$. Returning to the overview of the proof given in Subsection 2.1, we are now going to use the fact that edges with colors d_i and m_i have a common vertex in U_{m_i} in order to deduce the linear equation L_i. For the next claim recall that we have fixed elements $s_1 \in S_1, \ldots, s_{p-\ell} \in S_{p-\ell}$ and we consider an arbitrary set K_x as defined above.

Claim 3.4 *Let $1 \leq i \leq \ell$. Then the vertices $\{x_t \ : \ t \in [r-1] \setminus I_i\} \cup \{y_j \ : \ j \in W_i\} \cup y_{m_i}$ span an edge (of color d_i) if and only if there is an element $s_{d_i} \in S_{d_i}$ such that $\{s_j \ : \ j \in W_i\} \cup s_{m_i} \cup s_{d_i}$ satisfy equation L_i (defined in (1)).*

Proof: H contains an edge containing the vertices $\{x_t \ : \ t \in [r-1] \setminus I_i\} \cup \{y_j \ : \ j \in W_i\} \cup y_{m_i}$ if and only if (recall (4)) there is an $s_{d_i} \in S_{d_i}$ such that

$$y_{m_i} = b_i - M_{i,d_i} \cdot s_{d_i} - \sum_{j \in W_i} M_{i,j} \cdot y_j + \sum_{t \in [r-1] \setminus I_i} x_t \cdot (a_t^{m_i} + \sum_{j \in W_i} a_t^j \cdot M_{i,j}) \tag{6}$$

Using (5) this is equivalent to requiring that

$$\begin{aligned}
s_{m_i} + \sum_{t=1}^{r-1} a_t^{m_i} \cdot x_t &= b_i - M_{i,d_i} \cdot s_{d_i} - \sum_{j \in W_i} M_{i,j} \cdot (s_j + \sum_{t=1}^{r-1} a_t^j \cdot x_t) \\
&\quad + \sum_{t \in [r-1] \setminus I_i} x_t \cdot (a_t^{m_i} + \sum_{j \in W_i} a_t^j \cdot M_{i,j}) \\
&= b_i - M_{i,d_i} \cdot s_{d_i} - \sum_{j \in W_i} M_{i,j} \cdot s_j - \sum_{t=1}^{r-1} x_t \cdot \left(\sum_{j \in W_i} a_t^j \cdot M_{i,j} \right) \\
&\quad + \sum_{t \in [r-1] \setminus I_i} x_t \cdot (a_t^{m_i} + \sum_{j \in W_i} a_t^j \cdot M_{i,j}) \\
&= b_i - M_{i,d_i} \cdot s_{d_i} - \sum_{j \in W_i} M_{i,j} \cdot s_j - \sum_{t \in I_i} x_t \cdot \left(\sum_{j \in W_i} a_t^j \cdot M_{i,j} \right) \\
&\quad + \sum_{t \in [r-1] \setminus I_i} x_t \cdot a_t^{m_i}.
\end{aligned}$$

Using (2) in the last row above, we can write the above requirement as

$$s_{m_i} + \sum_{t=1}^{r-1} a_t^{m_i} \cdot x_t = b_i - M_{i,d_i} \cdot s_{d_i} - \sum_{j \in W_i} M_{i,j} \cdot s_j + \sum_{t=1}^{r-1} a_t^{m_i} \cdot x_t \ ,$$

or equivalently that

$$s_{m_i} + M_{i,d_i} \cdot s_{d_i} + \sum_{j \in W_i} M_{i,j} \cdot s_j = b_i \ ,$$

which is precisely equation L_i. ■

For the next two claims, let us recall that we assume that the last ℓ columns of M form a diagonal matrix. Therefore, a solution to $Mx = b$ is determined by the first $p - \ell$ elements of x.

Claim 3.5 *Suppose $s_1, \ldots, s_{p-\ell}$ determine a solution $s_1, \ldots, s_p$ to $Mx = b$. Then, any set K_x (defined above) spans a colored copy of K. In particular, for every solution $s_1, \ldots, s_p$ to $Mx = b$, H has n^{r-1} colored copies of K, in which the edge of color i is labeled with s_i.*

Proof: We claim that K_x spans a colored copy of K, where for every $1 \leq i \leq r-1$ vertex v_i of K is mapped to vertex x_i of H, and for every $1 \leq j \leq p - \ell$ vertex u_j of K is mapped to vertex y_j of H. To see that the above is a valid mapping of the colored edges of K to colored edges of H, we first note that the way we have defined H in (3) and the vertices $y_1, \ldots, y_{p-\ell}$ in (5), guarantees that for every $1 \leq j \leq p - \ell$ we have an edge with color i which contains the vertices $x_1, \ldots, x_{r-1}, y_j$. This is actually true even if $s_1, \ldots, s_{p-\ell}$ do not determine a solution.

As for edges with color $p-\ell+1 \leq d_i \leq p$, the fact that the vertices $\{x_t \ : \ t \in [r-1] \setminus I_i\} \cup \{y_j \ : \ j \in W_i\} \cup y_{m_i}$ span such an edge follows from Claim 3.4, because we assume that $s_1, \ldots, s_{p-\ell}$ determine a solution to $Mx = b$, so for every $1 \leq i \leq \ell$ there exists an element $s_{d_i} \in S_{d_i}$ as required by Claim 3.4. We thus conclude that $x_1, \ldots, x_{r-1}, y_1, \ldots, y_{p-\ell}$ span a colored copy of K. Finally, note that by the way we have defined H, the edge of K_x which is colored i is indeed labeled with the element $s_i \in S_i$. ■

Claim 3.6 *If $s_1, \ldots, s_{p-\ell}$ determine a solution to $Mx = b$, then the n^{r-1} colored copies of K spanned by the sets K_x (defined above) are edge disjoint.*

Proof: Let us consider two colored copies K_x and K_y for some $x \neq y$ (Claim 3.5 guarantees that K_x and K_y indeed span a colored copy of K). Clearly K_x and K_y cannot share edges with color $1 \leq i \leq p-\ell$, because the vertices of such edges within $V_1, \ldots, V_{r-1}$ are uniquely determined by the coordinates of x and y.

We now consider an edge of K_x with color $d_i \in \{p-\ell+1, \ldots, p\}$. Let $j_1 < j_2 < \ldots < j_{|W_i|}$ be the elements of W_i, and let B_i be the matrix defined in Claim 3.2. Recall that B_i satisfies the following[8]: (i) for $j \in [r-1] \setminus I_i$ we have $(B_i)_{j,j} = 1$ and $(B_i)_{j,t} = 0$ when $t \neq j$, and (ii) if $j \in I_i$ is the g^{th} element of I_i, then the j^{th} row of B_i is the vector a^{j_g} (where j_g is the g^{th} element of W_i). Let us also define an $r-1$ dimensional vector c as follows: for every $j \in [r-1] \setminus I_i$ we have $c_j = 0$, and for every $j \in I_i$, if j is the g^{th} element of I_i then $c_j = s_{j_g}$. The key observation now is that the vertices of the edge whose color is $d_i \in \{p-\ell+1, \ldots, p\}$ within the $r-1$ sets $\{V_j \ : \ j \in [r-1] \setminus I_i\} \cup \{U_j \ : \ j \in W_i\}$ are given by $B_i x + c$. More precisely, for every $j \in [r-1] \setminus I_i$ the vertex of the edge of color d_i within V_j is given by $(B_i x + c)_j$. Also, for every $j_g \in W_i$, if $j \in I_i$ is the g^{th} element of I_i, then the vertex of this edge within U_{j_g} is given by $(B_i x + c)_j$. Claim 3.2 asserts that B_i is non-singular, so we can conclude that the edges with color d_i of K_x and K_y can share at most $r-2$ of their $r-1$ vertices within the sets $\{V_j \ : \ j \in [r-1] \setminus I_i\} \cup \{U_j \ : \ j \in W_i\}$. So any pair of edges of color d_i can share at most $r-1$ vertices, and therefore K_x and K_y are edge disjoint [9]. ■

Claim 3.7 *If $S_1, \ldots, S_p$ contain at most T solutions to $Mx = b$ with $x_i \in S_i$ then H contains at most Tn^{r-1} colored copies of K.*

Proof: Recall that we assume that the last ℓ columns of M form a diagonal matrix. Therefore, the number of solutions T to $Mx = b$ is just the number of choices of $s_1 \in S_1, \ldots, s_{p-\ell} \in S_{p-\ell}$ that can be extended to a solution of $Mx = b$ by choosing appropriate values $s_{p-\ell+1} \in S_{p-\ell+1}, \ldots, s_p \in S_p$.

[8] We remark that when we have defined the matrices B_i in Claim 3.2 we did not "impose" the ordering of the rows that correspond to W_i as we do here, but this ordering, of course, does not affect the rank of B_i.

[9] We note that the way we have defined H does not (necessarily) guarantee that edges of the same color cannot share $r-1$ vertices. That is, edges of color i may share the vertex in the set U_{m_i} and $r-2$ of the $r-1$ vertices from the sets $\{V_j \ : \ j \in [r-1] \setminus I_i\} \cup \{U_j \ : \ j \in W_i\}$.

Therefore, it is enough to show that every colored copy of K in H is given by a choice of $r-1$ vertices $x_1 \in V_1, \ldots, x_{r-1} \in V_{r-1}$ and a choice of $p-\ell$ elements $s_1 \in S_1, \ldots, s_{p-\ell} \in S_{p-\ell}$ that determine a solution to $Mx = b$. So let us consider a colored copy of K in H. This copy must contain edges with the colors $1, \ldots, p-\ell$. By the way we have defined H this means that this copy must contain $r-1$ vertices $x_1 \in V_1, \ldots, x_{r-1} \in V_{r-1}$ as well as $p-\ell$ vertices $y_1 \in U_1, \ldots, y_{p-\ell} \in U_{p-\ell}$. Furthermore, for $1 \leq j \leq p-\ell$ we have

$$y_j = s_j + \sum_{t=1}^{r-1} a_t^j \cdot x_t \tag{7}$$

for some choice of $s_j \in S_j$. So the vertex set of such a copy is determined by the choice of $x_1, \ldots, x_{r-1}$ and $s_1, \ldots, s_{p-\ell}$. Note that the set of vertices is just the set K_x defined before Claim 3.4, for $x_1, \ldots, x_{r-1}$ and $s_1, \ldots, s_{p-\ell}$. Therefore, we can apply Claim 3.4 on this set of vertices.

So our goal now is to show that there are elements $s_{p-\ell+1}, \ldots, s_p$ which together with $s_1, \ldots, s_{p-\ell}$ form a solution of $Mx = b$. Consider any $1 \leq i \leq \ell$. As the vertices at hand span a colored copy of K they must span an edge with color d_i. This edge must[10] contain the vertices $\{x_t \ : \ t \in [r-1] \setminus I_i\} \cup \{y_j \ : \ j \in W_i\} \cup y_{m_i}$. But by Claim 3.4 if these vertices span an edge (of color d_i) then there is an element $s_{d_i} \in S_{d_i}$ such that $\{s_j \ : \ j \in W_i\} \cup s_{m_i} \cup s_{d_i}$ satisfy equation L_i. As this holds for every $1 \leq i \leq \ell$ we deduce that $s_1, \ldots, s_p$ satisfy $Mx = b$. ■

The proof of Lemma 2.4 now follows from Claims 3.3, 3.5, 3.6 and 3.7.

4 Testing Properties of Boolean Functions

Property testing algorithms, or testers for short, are fast randomized algorithms for distinguishing between objects satisfying a property and those that are "far" from satisfying it. For example, if the object we are interested in is a boolean function $f : D \mapsto \{0,1\}$, for some domain D, then we say that f is δ-far from satisfying a property $\mathcal{P}$ if one should edit the truth table of f in at least $\delta|D|$ places to get a function satisfying the property. As another example, if the object we are interested in is a subset S of some finite field $\mathbb{F}$, then we say that S is δ-far from satisfying a property $\mathcal{P}$ if one should add to or delete from S at least $\delta|\mathbb{F}|$ elements to get a set satisfying the property. As opposed to algorithms for (exactly) deciding if a given object satisfies a property, δ-testing algorithms, or testers for short, are only required to distinguish with high probability (say, 2/3) between objects satisfying the property and those that are δ-far from satisfying the property. The ultimate goal is to devise a testing algorithm whose running time depends only on the error parameter δ, and is independent of the size of the input. This area of research was first defined in [10, 28], where properties of boolean functions (such as linearity) were studied. It was further studied later in [15] in the context of combinatorial structures. See [12] for a recent survey and references.

[10] Because only vertices from this combination of r of the sets $V_1, \ldots, V_{r-1}, U_1, \ldots, U_{p-\ell}$ spans an edge with color d_i.

Our investigation thus far was about properties of combinatorial structures, but the main result has applications to the study of properties of boolean functions. One of the most interesting questions in the area of property testing is which properties have efficient testing algorithms, that is, algorithms whose running time is a function of the error parameter δ. Such problems were extensively studied in the context of testing properties of graphs, see, e.g., [3] and [15]. But there are no such general results on testing properties of boolean functions. In an attempt to find such a general result Bhattacharyya, Chen, Sudan and Xie [9], following an earlier result of Kaufman and Sudan [20], formulated a family of properties of boolean functions, and conjectured that any property that belongs to this family can be efficiently tested. As it turns out, if instead of looking at a boolean functions $f : \mathbb{F} \mapsto \{0, 1\}$, one considers[11] subsets $S \subseteq \mathbb{F}$, then the conjecture of [9] is equivalent to showing that for every set of homogenous linear equations $Mx = 0$, one can efficiently test the property of subsets S for the property of being $(M, 0)$-free. As the following theorem shows, an easy application of Theorem 2 gives that we can actually test even non-homogenous sets of linear equations. For more details on the relation of our main result to property testing the reader is referred to the conference version of this paper [32].

Theorem 5 *For every set of linear equations $Mx = b$ over $\mathbb{F}_n$, the property of being (M, b)-free has a constant time testing algorithm.*

Proof: Let $Mx = b$ be a set of ℓ linear equations in p unknowns. Definition 1.1 guarantees that if this set of equations has the removal property, then for every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, p)$, such that if S is δ-far from being (M, b)-free, then S contains at least $\epsilon n^{p-\ell}$ solutions to $Mx = b$. As $S \subseteq \mathbb{F}_n$ and $|\mathbb{F}_n| = n$, this means that if we pick a random vector $v \in S^{p-\ell}$, then with probability at least ϵ it satisfies $Mv = b$. Hence, it is enough for the δ-tester to pick $4/\epsilon$ such vectors v and check if any one of them satisfies $Mv = b$. If one of them satisfies the linear equations the tester rejects, otherwise it accepts. The tester clearly accepts with probability 1 any S that is (M, b)-free. Furthermore, if S is δ-far from being (M, b)-free, then the tester rejects S with probability at least $1 - (1 - \epsilon)^{4/\epsilon} > 2/3$, as needed. Finally, note that for any fixed set of linear equations, the running time of the algorithm is $O(1/\epsilon)$, where ϵ depends only on δ. ■

5 Concluding Remarks and Open Problems

5.1 The types of bounds we get

Our proof of the removal lemma for sets of linear equations applies the hypergraph removal lemma. As a consequence, we get extremely poor bounds relating ϵ and δ. Roughly speaking, the best current bounds for the graph removal lemma give that $\epsilon(\delta)$ grows like Tower$(1/\delta)$, that is, a tower

[11] By simply looking at the characteristic sets $S_f = \{x \in \mathbb{F} \ : \ f(x) = 1\}$.

of exponents of height $1/\delta$. For 3-uniform hypergraphs, the bounds are given by iterating the Tower function $1/\delta$ times. More generally, the bounds for r-uniform hypergraphs grow like the r^{th} function in the Ackermann hierarchy. Therefore, our bounds are also of this type. It seems very interesting to come up with a proof which would avoid using the removal lemma, and would thus supply tighter bounds.

5.2 On the possibility of improved bounds

Given the above discussion it is reasonable to ask for which sets of equations $Mx = b$ one can get a polynomial dependence between ϵ and δ. More precisely, which sets of linear equations $Mx = b$ have the property that if one should remove δn elements from $S \subseteq [n]$ in order to make it (M, b)-free then S contains at least $\epsilon n^{p-\ell}$ solutions to $Mx = b$, where $\epsilon = \epsilon(\delta) \geq \delta^C$ for some constant $C = C(M, b)$. Note that Theorem 2 guarantees that for any $\delta > 0$ there is an $\epsilon(\delta) > 0$ satisfying the above assertion. However, as we have mentioned in the previous subsection, this dependence is far from being polynomial. This problem seems to be a challenging open problem even for a single homogenous equation so let us focus on this case.

A solution to a linear equation $\sum a_i x_i = 0$ is non-trivial if all the x_i are distinct. For a linear equation L, let $r_L(n)$ denote the size of the largest subset of n which contains no non-trivial solution to L. Problems of this type were studied by Ruzsa [29], see also [31]. By applying an argument from [29] it can be shown that for certain equations satisfying $r_L(n) = n^{1-c}$ for a positive c, we have $\epsilon(\delta) \geq \delta^C$. However, characterizing the equations L satisfying $r_L(n) = n^{1-c}$ seems like a very hard problem, see [29]. Furthermore, we do not even know if all the linear equations for which $r_L(n) = n^{1-o(1)}$ do not have a polynomial dependence between δ and ϵ. For example, a special case for which we do not know if such a dependence exists is for the linear equation $x_1 + x_2 = x_3$ (for which $r_L(n) = \Theta(n)$). But for at least some of these linear equations, we can rule out such a polynomial dependence as the following example shows. Recall that $x + y = 2z$ if and only if x, z, y form a 3-term arithmetic progression. We call this progression trivial if $x = y = z$.

Proposition 5.1 *If L is the linear equation $x + y = 2z$ then we have $\epsilon(\delta) < \delta^{c \log 1/\delta}$. That is, for every $\delta > 0$ there is a set $S \subseteq [n]$ such that one should remove at least δn elements from S in order to destroy all the (non-trivial) solutions of L in S, and yet S contains only $\delta^{c \log 1/\delta} n^2$ solutions of L.*

Proof: Fix a $\delta > 0$ and let $n_0 = n_0(\delta)$ be large enough that for every $n \geq n_0$, every $X \subseteq [n]$ of size δn contains a non-trivial 3-term arithmetic progression. Roth's Theorem [29] states that such an n_0 exists. Therefore, for every $n \geq n_0$ and for every $X \subseteq [n]$ of size $2\delta n$ we have to remove at least δn elements from X in order to destroy all 3-term arithmetic progressions. Let $m = m(\delta)$ be the largest integer for which $[m]$ contains a subset of size $4\delta m$, containing no non-trivial 3-term arithmetic progressions. The well known construction of Behrend [6] shows that there are subsets of

$[n]$ of size $n/2^{c\sqrt{\log n}}$ containing no non-trivial 3-term arithmetic progressions. This means that

$$m = m(\delta) \geq (1/\delta)^{c\log(1/\delta)} \,. \tag{8}$$

for some absolute constant c. Let X be a subset of $[m]$ of size $4\delta m$, containing no non-trivial 3-term arithmetic progressions.

We are now ready to define the set S. For every $n \geq n_0$, let $S \subseteq [n]$ be the set of integers with the property that in their base $2m$ representation, the least significant element belongs to X. Then clearly $|S| = n \cdot \frac{|X|}{2m} = 2\delta n$ and so one should remove at least δn elements from S to destroy all 3-term arithmetic progressions. Since there is no carry when adding the least significant elements of $x_1, x_2 \in S$, we conclude that if $x_1, x_2, x_3 \in S$ form a 3-term arithmetic progression then the least significant characters of x_1, x_2, x_3 must also form a 3-term arithmetic progression. But as these characters belong to X we get that they must be identical (that is, this progression is trivial). Therefore, recalling (8), we infer that the number of 3-term arithmetic progressions in S is at most $n^2/m \leq \delta^{c\log 1/\delta}n^2$, thus completing the proof. ■

We note that it is not difficult to extend the argument in the above proof to any linear equation in which one variable is a convex combination of the others.

We finally mention that a particularly interesting investigation is when the field we are working in is $\mathbb{F}_2^n$. This is related (see Section 4) to testing properties of boolean functions. A fascinating open problem is whether there is *any* linear equation (over $\mathbb{F}_2^n$) for which the dependence between ϵ and δ is super-polynomial. The only result in this direction was obtained recently by Bhattacharyya and Xie [8] who showed that this relation is super-linear.

5.3 Non-monotone variants of Theorem 2

The property of sets being (M, b)-free is analogous to the graph property of being H-free. Note that both properties are monotone in the sense that removing elements from a set that is (M, b)-free results in an (M, b)-free set, just like removing edges from an H-free graph results in an H-free graph. A non-monotone variant of graphs being H-free is of course the property of being *induced* H-free. Alon et al. [1] obtained a removal lemma for the property of being induced H-free. More precisely, they have shown that for every graph fixed graph H and every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, H)$ such that if G is an n-vertex graph containing less than ϵn^h *induced* copies of H then one can add to or remove from G a set of at most δn^2 edges and thus obtain a graph with no induced copy of H. It is now natural to define the following "induced" variant of the property of being (M, b)-free

Definition 5.2 *Let $\mathbb{F}_n$ be the field of size n, let M be an $\ell \times p$ matrix over $\mathbb{F}_n$, let $b \in \mathbb{F}_n^\ell$ and let $I \subseteq [p]$. We say that S is (M, b, I)-free if there is no vector $v = \{v_1, \ldots, v_p\} \in \mathbb{F}_n^p$ satisfying $Mv = b$, where $v_i \in S$ iff $i \in I$.*

The set of linear equations $Mx = b$ has the induced removal property *if for every $\delta > 0$ there is an $\epsilon = \epsilon(\delta, p) > 0$ with the following property; Let $I \subseteq [p]$ and suppose $S \subseteq \mathbb{F}_n$ is such that there are at most $\epsilon n^{p-\ell}$ vectors $v = \{v_1, \ldots, v_p\} \in \mathbb{F}_n^p$ satisfying $Mv = b$, where $v_i \in S$ iff $i \in I$, then one can add to or remove from S at most δn elements to obtain an (M, b, I)-free set.*

Observe that one can think of the set I in the above definition as being analogous to the edge set of the graph H in the property of being induced H-free. Note also that we now allow to both add and delete elements from S in order to destroy all solutions.

Given the result of [1] mentioned above, it is natural to conjecture that one can strengthen Theorem 2 by proving the following

Conjecture 2 *Every set of linear equations over any field has the induced removal property.*

Of course, a positive answer to the above conjecture will result in testing algorithms for the non-monotone variants of the property of being (M, b)-free via the arguments in Section 4.

5.4 A removal lemma for infinitely many systems of equations

The contrapositive version of our main result says that if one should remove at least δn elements from $S \subseteq [n]$ in order to destroy all solutions of $Mx = b$ then S contains $f_{M,b}(\delta)n^{p-\ell}$ solutions to $Mx = b$, where $f_{M,b}(\delta) > 0$ for every $\delta > 0$. The "analogous" result for graphs (or hypergraphs) is that if one should remove at least δn^2 edges from a graph G in order to destroy all the copies of H then G contains $f_H(\delta)n^h$ copies of H, where h is the number of vertices of H and $f_H(\delta) > 0$ for every $\delta > 0$. The main result of [3] is an "infinite" version of the removal lemma for graphs, which states that if $\mathcal{H}$ is a (possibly infinite) set of graphs, and if one should remove at least δn^2 edges from G in order to destroy all the copies of all the graphs $H \in \mathcal{H}$ then for some $H \in \mathcal{H}$, whose size h satisfies $h \leq h(\delta)$, the graph G must contain at least $f_{\mathcal{H}}(\delta)n^h$ copies of H.

It seems natural to ask if there is a corresponding "infinite" removal lemma for sets of linear equations. More precisely, is it the case that for every (possibly infinite) set $\mathcal{M} = \{M_1x = b_1, M_2x = b_2, \ldots\}$ of sets of linear equations the following holds: if one should remove at least δn elements from $S \subseteq [n]$ in order to destroy all the solutions to all the sets of linear equations in $\mathcal{M}$, then for some set of linear equations $Mx = b \in \mathcal{M}$, with $p \leq p(\epsilon)$ unknowns, S contains at least $f_{\mathcal{M}}(\delta)n^{p-\ell}$ solutions to $Mx = b$.

5.5 A removal lemma over groups

Our removal lemma for sets of linear equations works over any field. For the special case of a single linear equation, Král', Serra and Vena [23] (following Green [18]) proved a removal lemma over any group. It is natural to ask if a similar removal lemma over groups, or even just abelian groups, also holds for sets of linear equations. See [34] for a related recent result.

5.6 The Bergelson-Host-Kra Conjecture

Green [18] used the regularity lemma for groups in order to resolve a conjecture of Bergelson, Host, Kra and Ruzsa [7], which stated that every $S \subseteq [n]$ of size δn contains at least $(\delta^3 - o(1))n$ 3-term arithmetic progressions with a common difference. The analogous statement for arithmetic progressions of length more than 4 was shown to be false in [7]. So the only case left open is whether any $S \subseteq [n]$ of size δn contains at least $(\delta^4 - o(1))n$ 4-term arithmetic progressions with a common difference. Part of the motivation of Green for raising Conjecture 1 was that it may help in resolving the case of the 4-term arithmetic progression. It seems very interesting to see if Theorem 2 can indeed help in resolving this conjecture.

Acknowledgements: I would like to thank Vojta Rödl, Benny Sudakov and Terry Tao for helpful discussions related to this paper, and Madhu Sudan for sending me a preliminary version of [9]. I would also like to thank Pablo Candela for his helpful comments on the paper and Tim Austin for communicating Conjecture 2 to me.

References

[1] N. Alon, E. Fischer, M. Krivelevich and M. Szegedy, Efficient testing of large graphs, Proc. of 40^{th} FOCS, New York, NY, IEEE (1999), 656–666. Also: Combinatorica 20 (2000), 451-476.

[2] N. Alon and A. Shapira, Testing Subgraphs in Directed Graphs, Journal of Computer and System Sciences, 69 (2004), 354-382.

[3] N. Alon and A. Shapira, Every monotone graph property is testable, SIAM J. on Computing, 38 (2008), 505-522.

[4] T. Austin, Private communication, 2008.

[5] T. Austin and T. Tao, On the testability and repair of hereditary hypergraph properties, Random Structures and Algorithms, to appear.

[6] F. A. Behrend, On sets of integers which contain no three terms in arithmetic progression, Proc. National Academy of Sciences USA 32 (1946), 331-332.

[7] V. Bergelson, B. Host, B. Kra and I.Z. Ruzsa, Multiple recurrence and nilsequences, Inventiones Mathematicae 160 (2005), 261-303.

[8] A. Bhattacharyya and N. Xie, Lower bounds for testing triangle freeness in boolean functions, Proc. of SODA 2010, to appear.

[9] A. Bhattacharyya, V. Chen, M. Sudan and N. Xie, Testing linear-invariant non-linear properties, Proc. of STACS 2009, 135-146.

[10] M. Blum, M. Luby and R. Rubinfeld, Self-testing/correcting with applications to numerical problems, JCSS 47 (1993), 549-595.

[11] P. Candela, On systems of linear equations and uniform hypergraphs, manuscript, 2008.

[12] A. Czumaj and C. Sohler, Sublinear-time algorithms, Bulletin of the EATCS, 89 (2006), 23-47.

[13] P. Frankl and V. Rödl, Extremal problems on set systems, Random Structures and Algorithms 20 (2002), 131-164.

[14] H. Furstenberg and Y. Katznelson, An ergodic Szemerédi theorem for commuting transformations, J. Analyse Math. 34 (1978), 275-291.

[15] O. Goldreich, S. Goldwasser and D. Ron, Property testing and its connection to learning and approximation, JACM 45(4): 653-750 (1998).

[16] T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, Ann. of Math. Volume 166, Number 3 (2007), 897-946.

[17] T. Gowers, Quasirandomness, counting and regularity for 3-uniform hypergraphs, Combinatorics, Probability and Computing, 15 (2006), 143-184.

[18] B. Green, A Szemerédi-type regularity lemma in abelian groups, GAFA 15 (2005), 340-376.

[19] Y. Ishigami, A simple regularization of hypergraphs, http://arxiv.org/abs/math/0612838.

[20] T. Kaufman and M. Sudan, Algebraic property testing: the role of invariance, Proc. of STOC 2008, 403-412.

[21] Y. Kohayakawa, B. Nagle, V. Rödl, M. Schacht and J. Skokan, The hypergraph regularity method and its applications, Proceedings of the National Academy of Sciences USA, 102(23): 8109-8113.

[22] J. Komlós and M. Simonovits, Szemerédi's Regularity Lemma and its applications in graph theory. In: *Combinatorics, Paul Erdös is Eighty*, Vol II (D. Miklós, V. T. Sós, T. Szönyi eds.), János Bolyai Math. Soc., Budapest (1996), 295–352.

[23] D. Král', O. Serra and L. Vena, A combinatorial proof of the removal lemma for groups, J. of Combinatortial Theory Ser. A, to appear.

[24] D. Král', O. Serra and L. Vena, A removal lemma for linear systems over finite fields, Jornadas de Matematica Discreta y algortimica 2008.

[25] B. Nagle, V. Rödl and M. Schacht, The counting lemma for regular k-uniform hypergraphs, Random Structures and Algorithms 28 (2006), 113-179.

[26] V. Rödl and J. Skokan, Regularity lemma for k-uniform hypergraphs, Random Structures and Algorithms 25 (2004), 1-42.

[27] K.F. Roth, On certain sets of integers, J. London Math. Soc. 28 (1953), 104-109.

[28] R. Rubinfeld and M. Sudan, Robust characterization of polynomials with applications to program testing, *SIAM J. on Computing* 25 (1996), 252–271.

[29] I. Z. Ruzsa, Solving a linear equation in a set of integers I, Acta Arithmetica 65 (1993), 259-282.

[30] I. Ruzsa and E. Szemerédi, Triple systems with no six points carrying three triangles, in Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai 18, Volume II, 939-945.

[31] A. Shapira, Behrend-type constructions for sets of linear equations, Acta Arithmetica, 122 (2006), 17-33.

[32] A. Shapira, Green's conjecture and testing linear-invariant properties, Proc. of STOC 2009, 159-166.

[33] A. Shapira and R. Yuster, Multigraphs (only) satisfy a weak triangle removal lemma, Electronic Journal of Combinatorics, 16 (2009), N11.

[34] B. Szegedy, The symmetry preserving removal lemma, arXiv:0809.2626v1.

[35] E. Szemerédi, Integer sets containing no k elements in arithmetic progression, Acta Arith. 27 (1975), 299-345.

[36] E. Szemerédi, Regular partitions of graphs, In: *Proc. Colloque Inter. CNRS* (J. C. Bermond, J. C. Fournier, M. Las Vergnas and D. Sotteau, eds.), 1978, 399–401.

[37] T. Tao, A variant of the hypergraph removal lemma, J. Combin. Theory, Ser. A 113 (2006), 1257-1280.