# A New Multilayered PCP and the Hardness of Hypergraph Vertex Cover

Irit Dinur * Venkatesan Guruswami† Subhash Khot‡ Oded Regev §

November 8, 2002

**Abstract**

Given a $k$-uniform hyper-graph, the E$k$-Vertex-Cover problem is to find the minimum-size subset of vertices that 'hits' every hyper-edge. We present a new multilayered PCP construction that extends the Raz verifier. This enables us to prove that E$k$-Vertex-Cover is NP-hard to approximate within factor $(k-1-\varepsilon)$ for any $k \geq 3$ and any $\varepsilon > 0$. The result is essentially tight as this problem can be easily approximated within factor $k$. Our construction makes use of the biased Long-Code and is analyzed using combinatorial properties of $s$-wise $t$-intersecting families of subsets.

## 1 Introduction

A $k$-uniform hypergraph $H = (V, E)$ consists of a set of vertices $V$ and a collection $E$ of $k$-element subsets of $V$ called hyperedges. A *vertex cover* of $H$ is a subset $S \subseteq V$ such that every hyperedge in $E$ intersects $S$, i.e., $e \cap S \neq \emptyset$ for each $e \in E$. An *independent set* in $G$ is a subset whose complement is a vertex cover, or in other words a subset of vertices that contains no hyperedge entirely within it. The E$k$-Vertex-Cover problem is the problem of finding a minimum size vertex cover in a $k$-uniform hypergraph. This problem is alternatively called the minimum hitting set problem with sets of size $k$ (and is equivalent to the set cover problem where each element of the universe occurs in exactly $k$ sets).

The E$k$-Vertex-Cover problem is a fundamental NP-hard optimization problem which arises in numerous settings. For $k = 2$, it is just the famous vertex cover problem on graphs. Owing to its NP-hardness, one is interested in how well it can be approximated in polynomial time. A very simple algorithm that is invariably taught in a typical undergraduate algorithms class is the following: greedily pick a maximal set of pairwise disjoint hyperedges and then include all vertices in the chosen hyperedges in the vertex cover. It is easy to show that this gives a factor $k$ approximation algorithm for E$k$-Vertex-Cover. State of the art techniques yield only a tiny improvement, achieving a $k - o(1)$ approximation ratio [12]. This raises the question whether achieving an approximation factor of $k - \varepsilon$ for any constant $\varepsilon > 0$ could be NP-hard. This question is one of the central open

*NEC Research Institute, Princeton, NJ. E-Mail: `iritd@research.nj.nec.com`.

†Department of Computer Science, University of Washington, Seattle, WA 98195. E-Mail: `venkat@cs.washington.edu`. Work done while at UC Berkeley as a Miller Research Fellow.

‡Department of Computer Science, Princeton University, Princeton, NJ 08544. E-Mail: `khot@cs.princeton.edu`

§Institute for Advanced Study, Princeton, NJ. E-Mail: `odedr@ias.edu`. Research supported by NSF grant CCR-9987845.

questions in the field of approximation algorithms (for instance, it figures as the second entry in the list of open problems in the final chapter of Vazirani's recent book [24]).

In this paper, we prove a nearly tight hardness result for E$k$-Vertex-Cover. Specifically, we prove that E$k$-Vertex-Cover is indeed NP-hard to approximate within factor $(k-1-\varepsilon)$ for any $\varepsilon > 0$, thus explaining why no efficient algorithm with performance guarantee much better than $k$ has been found.

### Previous Hardness Results

The vertex-cover problem on hypergraphs where the size of the hyperedges is unbounded is no other than Set-Cover. For this problem there is a $\ln n$ approximation algorithm [20, 18], and a matching $(1-o(1))\ln n$ hardness result due to Feige [8]. The first explicit hardness result shown for E$k$-Vertex-Cover was due to Trevisan [23] who considered the approximability of bounded degree instances of several combinatorial problems, and specifically showed an inapproximability factor of $k^{1/19}$ for E$k$-Vertex-Cover. Holmerin [16] showed that E4-Vertex-Cover is NP-hard to approximate within $(2-\varepsilon)$, and more recently [17] that E$k$-Vertex-Cover is NP-hard to approximate within a factor of $k^{1-\varepsilon}$, and also that it is NP-hard to approximate E3-Vertex-Cover within factor $(3/2-\varepsilon)$. Goldreich [10] showed a direct 'FGLSS'-type [9] reduction (involving no use of the long-code, a crucial component in most recent PCP constructions) attaining a hardness factor of $(2-\varepsilon)$ for E$k$-Vertex-Cover for some constant $k$.

Somewhat surprisingly, more recently Dinur, Guruswami and Khot gave a fairly *simple* proof of an $\alpha \cdot k$ hardness result for E$k$-Vertex-Cover, (for some $\alpha > \frac{1}{3}$). The proof takes a combinatorial view of Holmerin's construction and instead of Fourier analysis uses some properties concerning intersecting families of finite sets. The authors also give a more complicated reduction that shows a factor $(k-3-\varepsilon)$ hardness for E$k$-Vertex-Cover. The crucial impetus for that work came from the recent result of Dinur and Safra [7] on the hardness of approximating vertex cover (on graphs), and as in [7] the notion of biased long codes and some extremal combinatorics relating to intersecting families of sets play an important role. In addition to ideas from [7], the factor $(k-3-\varepsilon)$ hardness result also exploits the notion of covering complexity introduced by Guruswami, Håstad and Sudan [11]. Both the $\alpha \cdot k$ and the $k-3-\varepsilon$ results have not been published (an ECCC manuscript exists, [5]) since they have been subsumed by the work presented herein.

### Our result and techniques

In this paper we improve upon all the above hardness results by proving a factor $(k-1-\varepsilon)$ inapproximability result for E$k$-Vertex-Cover. Already for $k=3$, this is an improvement from $1.5-\varepsilon$ to $2-\varepsilon$. Extending our result from $k-1-\varepsilon$ to $k-\varepsilon$ appears highly non-trivial and in particular would imply a factor $2-\varepsilon$ hardness for vertex-cover on graphs, a problem that is notoriously difficult. While our proof shares some of the extremal combinatorics flavor of [7] and [5], it draws its strength mainly from a new multilayered outer verifier system for NP languages. This multilayered system is constructed using the Raz verifier [21] as a building block.

The Raz verifier, which serves as the starting point or "outer verifier" in most if not all recent hardness results, can be described as follows. There are two sets of (non-Boolean) variables $Y$ and $Z$, and for certain pairs of $y \in Y$ and $z \in Z$, a constraint $\pi_{y \to z}$. The constraints are projections, i.e., for each assignment to $y$ there exists exactly one assignment to $z$ such that the constraint $\pi_{y \to z}$ is satisfied. The goal is to find an assignment $A$ to the variables so that a maximum number of constraints $\pi_{y \to z}$ are satisfied, i.e., have the property $\pi_{y \to z}(A(y)) = A(z)$. The PCP Theorem [2, 1] along with the Parallel Repetition Theorem [21] imply that for any $\varepsilon > 0$ it is NP-hard to distinguish between the case where all the constraints can be satisfied and the case where no more than $\varepsilon$ of the constraints can be satisfied.

In [5], the $\alpha \cdot k$ hardness result is obtained by replacing every $Y$ variable by a whole block of vertices (representing its Long-Code). Hyperedges connect $y_1$-vertices to $y_2$-vertices only if there is some $z \in Z$ such that $\pi_{y_1 \to z}, \pi_{y_2 \to z}$ are constraints in the system. This construction has an inherent symmetry between blocks which deteriorates the projection property of the constraints, limiting the hardness factor one can prove to at most $k/2$.

Another way of reducing the Raz verifier to E$k$-Vertex-Cover is to have for each variable in $Y$ and $Z$, a block of vertices (representing its Long-Code). Each constraint $\pi_{y \to z}$ can be emulated by a set of hyperedges, where each hyperedge consists of both $y$-vertices and $z$-vertices. The hyperedges can be chosen so that if the initial PCP instance was satisfiable, then taking a certain $1/k$ of the vertices in each block will be a vertex-cover. However, this reduction has a basic flaw: the underlying constraint graph, being bipartite with parts $Y$ and $Z$, has a vertex cover of size at most one half of the number of vertices. Taking all the vertices of, say, the $Z$ variables will be a vertex cover for the hypergraph regardless of whether or not the initial PCP instance was satisfiable. This, once more, limits the gap to no more than $k/2$.

Therefore, in order to overcome the limit of $k/2$, we need a version of the Raz verifier in which the constraints are projections and in which unsatisfiable PCP instances would be reduced to hypergraphs where the minimum vertex cover contains *almost all* the vertices.

We achieve this goal by presenting a new, multilayered PCP. In the usual Raz verifier we have two 'layers', the first containing the $Y$ variables and the second containing the $Z$ variables. In the multilayered PCP, we have $l$ layers containing variables $X_1, X_2, \ldots, X_l$ respectively. Between every pair of layers $i_1$ and $i_2$, we have a set of projection constraints that represent an instance of the Raz verifier. In the multilayered PCP, it is NP-hard to distinguish between the case where there exists an assignment that satisfies *all* the constraints (between any two layers) and the case where for *every* $i_1, i_2$ it is impossible to satisfy more than $\varepsilon$ of the constraints between $X_{i_1}$ and $X_{i_2}$.

In addition, we prove that the underlying constraint graph no longer has the obstacle of having a small vertex cover (hence a large independent set). Indeed we show that the multilayered PCP has a certain 'weak-density' property: for any set containing an $\varepsilon$ fraction of the variables there are many constraints between variables of the set. This guarantees that "fake" independent sets in the hypergraph (i.e., independent sets that occur because there are no constraints between the variables of the set) contain at most $\varepsilon$ of the vertices. Hence, the minimum vertex cover must contain vertices in almost all of the blocks.

We view the construction of the multilayered PCP as a central contribution of our paper, and believe that it could be a powerful tool to reduce from in other hardness of approximation results as well. In fact, our multilayered construction has already been used in obtaining strong hardness results for coloring 3-uniform hypergraphs [6, 19] (namely the hardness of coloring a 2-colorable 3-uniform hypergraph using an arbitrary constant number of colors), a problem for which no non-trivial inapproximability results are known using other techniques. We anticipate that this new outer verifier will also find other applications besides the ones in this paper and in [6, 19].

**The Biased Long-Code.** Our hypergraph construction relies on the Long-Code that was introduced in [3], and more specifically, on the biased Long-Code defined in [7]. Thus, each PCP variable is represented by a block of vertices, one for each 'bit' of the biased Long-Code. More specifically, in $x$'s block we have one vertex for each subset of $R$, where $R$ is the set of assignments for the variable $x$. However, rather than taking all vertices in a block with equal weight, we attach weights to the vertices according to the $p$-biased Long-Code. The weight of a subset $F$ is set to $p^{|F|}(1-p)^{|R \setminus F|}$, highlighting subsets of cardinality $p \cdot |R|$. Thus we actually construct a weighted hypergraph which can then be easily translated, by appropriate duplication of vertices, to a non-weighted one (see, e.g., [7]).

The vertex cover in the hypergraph is shown to have relative size of either $1-p$ in the good case or almost 1 in the bad case. Choosing large $p = 1 - \frac{1}{k-1-\varepsilon}$, yields the desired gap of $\frac{1}{1-p} \approx k-1-\varepsilon$ between the good and bad cases. The reduction uses the following property: *a family of subsets of a set $R$, where each subset has size $p|R|$, either contains very few subsets, or it contains some $k-1$ subsets whose common intersection is very small.* We will later show that this property holds for $p < 1 - \frac{1}{k-1}$ and therefore we obtain a gap of $k-1-\varepsilon$. As can be seen, this property does not hold for $p > 1 - \frac{1}{k-1}$ and therefore one cannot improve the $k-1-\varepsilon$ result by simply increasing $p$.

**Location of the gap.** All our hardness results have the gap between sizes of the vertex cover at the "strongest" location. Specifically, to prove a factor $(k-1-\varepsilon)$ hardness we show that it is hard to distinguish between $k$-uniform hypergraphs that have a vertex cover of weight $\frac{1}{k-1} + \varepsilon$ from those whose minimum vertex cover has weight at least $(1-\varepsilon)$. This result is stronger than a gap of about $(k-1)$ achieved, for example, between vertex covers of weight $\frac{1}{(k-1)^2}$ and $\frac{1}{k-1}$. In fact, by adding dummy vertices, our result implies that for any $c < 1$ it is NP-hard to distinguish between a graph whose vertex-cover has weight $c$ and $(\frac{c}{k-1} + \varepsilon)$. Put another way, our result shows that for $k$-uniform hypergraphs, for $k \geq 3$, there is a fixed $\alpha$ such that for arbitrarily small $\varepsilon > 0$, it is NP-hard to find an independent set consisting of a fraction $\varepsilon$ of the vertices even if the hypergraph is promised to contain an independent set comprising a fraction $\alpha$ of vertices. We remark that such a result is not known for graphs and seems out of reach of current techniques. (The recent 1.36 hardness result for vertex cover on graphs due to Dinur and Safra [7], for example, shows that it is NP-hard to distinguish between cases when the graph has an independent set of size $0.38 \cdot n$ and when no independent set has more than $0.16 \cdot n$ vertices.)

**Organization.** We begin in Section 2 by developing the machinery from extremal combinatorics concerning intersecting families of sets that will play a crucial role in our proof. In Section 3 we present the multilayered PCP construction. In Section 4 we present our reduction to a gap version of E$k$-Vertex-Cover which allows us to prove a factor $(k-1-\varepsilon)$ inapproximability result for this problem.

## 2 Intersecting Families

In this section we describe certain properties of $s$-wise $t$-intersecting families. For a comprehensive survey, see [13]. Denote $[n] = \{0, 1, \ldots, n-1\}$ and $2^{[n]} = \{F \mid F \subseteq [n]\}$.

**Definition 2.1** *A family $\mathcal{F} \subseteq 2^{[n]}$ is called $s$-*wise t-intersecting* if $|F_1 \cap \ldots \cap F_s| \geq t$ for any $s$ sets $F_1, \ldots, F_s \in \mathcal{F}$.*

We are interested in bounding the size of such families, and for this purpose it is useful to introduce the notion of a left-shifted family. Performing an $(i,j)$-shift on a family consists of replacing the element $j$ with the element $i$ in all sets $F \in \mathcal{F}$ such that $j \in F$, $i \notin F$ and $(F \setminus \{j\}) \cup \{i\} \notin \mathcal{F}$. A left-shifted family is a family which is invariant with respect to $(i,j)$-shifts for any $1 \leq i < j \leq n$. For any family $\mathcal{F}$, by iterating the $(i,j)$-shift for all $1 \leq i < j \leq n$ we eventually get a left-shifted family which we denote by $S(\mathcal{F})$. The following simple lemma summarizes the properties of the left-shift operation (see, e.g., [13], p. 1298, Lemma 4.2):

**Lemma 2.2** *For any family $\mathcal{F} \subseteq 2^{[n]}$, there exists a one-to-one and onto mapping $\tau$ from $\mathcal{F}$ to $S(\mathcal{F})$ such that $|F| = |\tau(F)|$ for every $F \in \mathcal{F}$. In other words, left-shifting a family maintains its size and the size of the sets in the family. Moreover, if $\mathcal{F}$ is an $s$-wise $t$-intersecting family then so is $S(\mathcal{F})$.* ■

The next lemma states that a subset $F$ in a left-shifted family, cannot be 'sparse' on all of its prefixes $F \cap [t + js]$, $\forall j \geq 0$.

**Lemma 2.3 ([13], p. 1311, Lemma 8.3)** *Let $\mathcal{F}$ be a left-shifted $s$-wise $t$-intersecting family. Then, for every $F \in \mathcal{F}$, there exists a $j \geq 0$ with $|F \cap [t + sj]| \geq t + (s - 1)j$.* ■

**Definition 2.4** *For a 'bias parameter' $0 < p < 1$, and a ground set $R$, the weight of a set $F \subseteq R$ is*

$$\mu_p^R(F) \stackrel{def}{=} p^{|F|} \cdot (1 - p)^{|R \setminus F|}$$

*When $R$ is clear from the context we write $\mu_p$ for $\mu_p^R$. The weight of a family $\mathcal{F} \subseteq 2^R$ is $\mu_p(\mathcal{F}) = \sum_{F \in \mathcal{F}} \mu_p(F)$.*

The weight of a subset is precisely the probability of obtaining this subset when one picks every element in $R$ independently with probability $p$.

The following is the main lemma of this section. It shows that for any $p < \frac{s-1}{s}$, a family of non-negligible $\mu_p$-weight (i.e., $\mu_p(\mathcal{F}) \geq \varepsilon$) cannot be $s$-wise $t$-intersecting for large enough $t$.

**Lemma 2.5** *For any $\varepsilon, s, p$ with $p < \frac{s-1}{s}$, there exists a $t = t(\varepsilon, s, p)$ such that for any $s$-wise $t$-intersecting family $\mathcal{F} \subseteq 2^{[n]}$, $\mu_p(\mathcal{F}) < \varepsilon$.*

**Proof:** The proof follows from Lemma 2.3 (see [13], p. 1311, Theorem 8.4). Let $\mathcal{F}$ be an $s$-wise $t$-intersecting family where $t$ will be determined later. According to Lemma 2.2, $S(\mathcal{F})$ is also $s$-wise $t$-intersecting and $\mu_p(S(\mathcal{F})) = \mu_p(\mathcal{F})$. By Lemma 2.3, for every $F \in S(\mathcal{F})$, there exists a $j \geq 0$ such that $|F \cap [t + sj]| \geq t + (s - 1)j$. We bound the size of the family $S(\mathcal{F})$ by the probability that such a $j$ exists for a random set chosen according to the distribution $\mu_p$.

Let $\delta = \frac{s-1}{s} - p$. Then, for any $j \geq 0$,

$$\begin{aligned}
\Pr[\ |F \cap [t + sj]| \geq t + (s - 1)j\ ] &\leq \Pr[\ |F \cap [t + sj]| - p(t + sj) \geq \delta(t + sj)\ ] \\
&\leq e^{-2(t+sj)\delta^2}
\end{aligned}$$

by the Chernoff bound [4]. Summing over all $j \geq 0$ we get:

$$\mu_p(S(\mathcal{F})) \leq \sum_{j \geq 0} e^{-2(t+sj)\delta^2} = e^{-2t\delta^2} / (1 - e^{-2s\delta^2})$$

which is smaller than $\varepsilon$ for large enough $t$. ■

# 3 The Multilayered PCP

## 3.1 Starting Point - The PCP Theorem and the Parallel Repetition Theorem

As is the case with many inapproximability results (e.g., [3], [14], [15], [22]), we begin our reduction from the Raz verifier described next. Let $\Psi$ be a PCP where the variables are of two types, denoted $Y$ and $Z$. Let $R_Y$ denote the range of the $Y$-variables and $R_Z$ the range of the $Z$-variables, where

$|R_Z| \leq |R_Y|$[1]. Assume each constraint $\pi \in \Psi$ depends on exactly one $y \in Y$ and one $z \in Z$, furthermore, for every value $a_y \in R_Y$ assigned to $y$ there is exactly one value $a_z \in R_Z$ to $z$ such that the constraint $\pi$ is satisfied. Therefore, we can write each constraint $\pi \in \Psi$ as a function from $R_Y$ to $R_Z$, and use notation $\pi_{y \to z} : R_Y \to R_Z$. Furthermore, we assume that the underlying constraint graph is bi-regular, i.e., every $Y$-variable appears in the same number of constraints in $\Psi$, and every $Z$-variable appears in the same number of constraints in $\Psi$.

The following theorem follows by combining the PCP Theorem with Raz's Parallel Repetition Theorem. The PCP given by this theorem will be called the Raz's verifier henceforth.

**Theorem 3.1 (PCP Theorem [1, 2] + Raz's Parallel Repetition Theorem [21])** *Let $\Psi$ be as above. There exists a universal constant $\gamma > 0$ such that for every (large enough) constant $|R_Y|$ it is NP-hard to distinguish between the following two cases:*

- **Yes :** *There is an assignment $A : Y \to R_Y$, $A : Z \to R_Z$ such that all $\pi \in \Psi$ are satisfied by $A$, i.e., $\forall \pi_{y \to z} \in \Psi$, $\pi_{y \to z}(A(y)) = A(z)$.*
- **No :** *No assignment can satisfy more than a fraction $\frac{1}{|R_Y|^\gamma}$ of the constraints in $\Psi$.* ■

As discussed in the introduction, a natural approach to build a hypergraph from the PCP $\Psi$ is to have a block of vertices for every variable $y$ or $z$ and define hyperedges of the hypergraph so as to enforce the constraints $\pi_{y \to z}$. For every constraint $\pi_{y \to z}$, there will be hyperedges containing vertices from the block of $y$ and the block of $z$. However, this approach is limited by the fact that the graph underlying the PCP has a small vertex cover. Since each hyperedge contains vertices from both the $Y$ and $Z$ 'sides', the subset of all vertices on the $Y$ (resp. $Z$) 'side', already covers all of the hyperedges regardless of whether the initial PCP system was satisfiable or not.

This difficulty motivates our construction of a multilayered PCP where we have many types of variables (rather than only $Y$ and $Z$) and the resulting hypergraph is *multipartite*. The multilayered PCP is able to maintain the properties of Theorem 3.1 between *every* pair of layers. Moreover, the underlying constraint graph has a special 'weak-density' property that roughly guarantees it will have only tiny independent sets (thus any vertex cover for it must contain almost all of the vertices).

## 3.2 Layering the Variables

Let $l, R > 0$. Let us begin by defining a $l$-layered PCP. In an $l$-layered PCP there are $l$ sets of variables denoted by $X_1, \ldots, X_l$. The range of variables in $X_i$ is denoted $R_i$, with $|R_i| = R^{O(l)}$. For every $1 \leq i < j \leq l$ there is a set of constraints $\Phi_{ij}$ where each constraint $\pi \in \Phi_{ij}$ depends on exactly one $x \in X_i$ and one $x' \in X_j$. For any two variables we denote by $\pi_{x \to x'}$ the constraint between them if such a constraint exists. Moreover, the constraints in $\Phi_{ij}$ are projections from $x$ to $x'$, that is, for every assignment to $x$ there is exactly one assignment to $x'$ such that the constraint is satisfied.

In addition, as mentioned in the introduction, we would like to show a certain 'weak-density' property of our multilayered PCP:

**Definition 3.2** *An $l$-layered PCP is said to be* weakly-dense *if for any $\delta > 0$, given $m \geq \lceil \frac{2}{\delta} \rceil$ layers $i_1 < \ldots < i_m$ and given any sets $S_j \subseteq X_{i_j}$ for $j \in [m]$ such that $S_j \geq \delta |X_{i_j}|$, there always exist two sets $S_j$ and $S_{j'}$ such that the number of constraints between them is at least a $\frac{\delta^2}{4}$ fraction of the constraints between the layers $X_{i_j}$ and $X_{i_{j'}}$.*

[1]Readers familiar with the Raz verifier may prefer to think concretely of $R_Y = [7^u]$ and $R_Z = [2^u]$ for some number $u$ of repetitions.

**Theorem 3.3** *There exists a universal constant $\gamma > 0$, such that for any parameters $l, R$, there is a weakly-dense $l$-layered PCP $\cup\Phi_{ij}$ such that it is NP-hard to distinguish between the following two cases:*

- **Yes :** *There exists an assignment that satisfies all the constraints.*
- **No :** *For every $i < j$, not more than $1/R^\gamma$ of the constraints in $\Phi_{ij}$ can be satisfied by an assignment.*

**Proof:** Let $\Psi$ be a constraint-system as in Theorem 3.1. We construct $\Phi = \cup\Phi_{ij}$ as follows. The variables $X_i$ of layer $i \in [l]$ are the elements of the set $Z^i Y^{l-i}$, i.e., all $l$-tuples where the first $i$ elements are $Z$ variables and the last $l-i$ elements are $Y$ variables. The variables in layer $i$ have assignments from the set $R_i = R_Z^i R_Y^{l-i}$ corresponding to an assignment to each variable of $\Psi$ in the $l$-tuple. It is easy to see that $|R_i| \le R^{O(l)}$ for any $i \in [l]$ and that the total number of variables is no more than $|\Psi|^{O(l)}$. For any $1 \le i < j \le l$ we define the constraints in $\Phi_{ij}$ as follows. A constraint exists between a variable $x_i \in X_i$ and a variable $x_j \in X_j$ if they contain the same $\Psi$ variables in the first $i$ and the last $l-j$ elements of their $l$-tuples. Moreover, for any $i < k \le j$ there should be a constraint in $\Psi$ between $x_{i,k}$ and $x_{j,k}$. More formally,

$$\Phi_{ij} = \left\{ \pi_{x_i,x_j} \;\middle|\; x_i \in X_i, x_j \in X_j, \begin{array}{ll} \forall k \in [l] \setminus \{i+1,\dots,j\}, & x_{i,k} = x_{j,k} \\ \forall k \in \{i+1,\dots,j\}, & \pi_{x_{i,k}\to x_{j,k}} \in \Psi \end{array} \right\}.$$

As promised, the constraints $\pi_{x_{i,k}\to x_{j,k}}$ are projections. Given an assignment $a \in Z^i Y^{l-i}$ to $x_i$, we define the consistent assignment $b \in Z^j Y^{l-j}$ to $x_j$ as $b_k = \pi_{x_{i,k}\to x_{j,k}}(a_k)$ for $k \in \{i+1,\dots,j\}$ and $b_k = a_k$ for all other $k$.

The completeness of $\Phi$ follows easily from the completeness of $\Psi$. That is, assume we are given an assignment $A : Y \cup Z \to R_Y \cup R_Z$ that satisfies all the constraints of $\Psi$. Then, the assignment $B : \bigcup X_i \to \bigcup R_i$ defined by $B(x_1 \dots x_l) = (A(x_1) \dots A(x_l))$ is a satisfying assignment.

For the soundness part, assume that there exist two layers $i < j$ and an assignment $B$ that satisfies more than a $1/R^\gamma$ fraction of the constraints in $\Phi_{ij}$. We partition $X_i$ into classes such that two variables in $X_i$ are in the same class iff they are identical except possibly on coordinate $j$. The variables in $X_j$ are also partitioned according to coordinate $j$. Since more than $1/R^\gamma$ of the constraints in $\Phi_{ij}$ are satisfied, it must be the case that there exist a class $x_{i,1},\dots,x_{i,j-1},x_{i,j+1},\dots,x_{i,l}$ in the partition of $X_i$ and a class $x_{j,1},\dots,x_{j,j-1},x_{j,j+1},\dots,x_{j,l}$ in the partition of $X_j$ between which there exist constraints and the fraction of satisfied constraints is more than $1/R^\gamma$. We define an assignment to $\Psi$ as $A(y) = (B(x_{i,1},\dots,x_{i,j-1},y,x_{i,j+1},\dots,x_{i,l}))_j$ for $y \in Y$ and as $A(z) = (B(x_{j,1},\dots,x_{j,j-1},z,x_{j,j+1},\dots,x_{j,l}))_j$ for $z \in Z$. Notice that there is a one-to-one and onto correspondence between the constraints in $\Psi$ and the constraints between the two chosen classes in $\Phi$. Moreover, if the constraint in $\Phi$ is satisfied, then the constraint in $\Psi$ is also satisfied. Therefore, $A$ is an assignment to $\Psi$ that satisfies more than $1/R^\gamma$ of the constraints.

To prove that this multilayered PCP is *weakly-dense*, we recall the bi-regularity property mentioned above, i.e., each variable $y \in Y$ appears in the same number of constraints and also each $z \in Z$ appears in the same number of constraints. Therefore, the distribution obtained by uniformly choosing a variable $y \in Y$ and then uniformly choosing one of the variables in $z \in Z$ with which it has a constraint is a uniform distribution on $Z$.

Take any $m = \lceil \frac{2}{\delta} \rceil$ layers $i_1 < \ldots < i_m$ and sets $S_j \subseteq X_{i_j}$ for $j \in [m]$ such that $S_j \ge \delta |X_{i_j}|$. Consider a random walk beginning from a uniformly chosen variable $x_1 \in X_1$ and proceeding to a variable $x_2 \in X_2$ chosen uniformly among the variables with which $x_1$ has a constraint. The random walk continues in a similar way to a variable $x_3 \in X_3$ chosen uniformly among the variables with which $x_2$ has a constraint and so on up to a variable in $X_l$. Denote by $E_j$ the indicator variable of the event that the random walk hits an $S_j$ variable when in layer $X_{i_j}$. From the uniformity of $\Psi$ it follows that for every $j$, $\Pr[E_j] \geq \delta$. Moreover, using the inclusion-exclusion principle, we get:

$$1 \geq \Pr[\bigvee E_j] \geq \sum_j \Pr[E_j] - \sum_{j<k} \Pr[E_j \wedge E_k]$$
$$\geq \lceil \frac{2}{\delta} \rceil \cdot \delta - \binom{m}{2} \max_{j<k} \Pr[E_j \wedge E_k] \geq 2 - \binom{m}{2} \max_{j<k} \Pr[E_j \wedge E_k]$$

which implies

$$\max_{j<k} \Pr[E_j \wedge E_k] \geq 1/\binom{m}{2} \geq \frac{\delta^2}{4}$$

Fix $j$ and $k$ such that $\Pr[E_j \wedge E_k] \geq \frac{\delta^2}{4}$ and consider a shorter random walk beginning from a random variable in $X_{i_j}$ and proceeding to the next layer and so on until hitting layer $i_k$. Since $E_j$ is uniform on $X_{i_j}$ we still have that $\Pr[E_j \wedge E_k] \geq \frac{\delta^2}{4}$ where the probability is taken over the random walks between $X_{i_j}$ and $X_{i_k}$. Also, notice that there is a one-to-one and onto mapping from the set of all random walks between $X_{i_j}$ and $X_{i_k}$ to the set $\Phi_{i_j,i_k}$. Therefore, $\frac{\delta^2}{4}$ of the constraints between $X_{i_j}$ and $X_{i_k}$ are between $S_j$ and $S_k$. ■

## 4 The Hypergraph Construction

**Theorem 4.1 (Main Theorem)** *For any $k \geq 3$ it is NP-hard to approximate the vertex-cover on a $k$-uniform hypergraph within any constant factor less than $k-1$.*

**Proof:** Fix $k \geq 3$ and arbitrarily small $\varepsilon > 0$. Define $p = 1 - \frac{1}{k-1-\varepsilon}$. Let $\Phi$ be a PCP instance with layers $X_1, \ldots, X_l$, as described in Theorem 3.3, with parameters $l = 32\varepsilon^{-2}$ and $R$ large enough to be chosen later. We present a construction of a $k$-uniform hypergraph $G = (V, E)$. We use the Long Code introduced by Bellare et al. [3]. A Long Code over domain $R$ has one bit for every subset $v \subseteq R$. An encoding of element $x \in R$ assigns bit-value 1 to the sets $v$ s.t. $x \in v$ and assigns 0 to the sets which do not contain $x$. In the following, the bits in the Long Code will be vertices of the hypergraph. The vertices that correspond to a bit-value 0 are (supposedly) the vertices of a Vertex Cover.

**Vertices.** For each variable $x$ in layer $X_i$ we construct a block of vertices $V[x]$. This block contains a vertex for each subset of $R_{X_i}$. Throughout this section we slightly abuse notation by writing a vertex rather than the set it represents. The weight of the vertices inside the block $V[x]$ is according to $\mu_p^{R_{X_i}}$, i.e. the weight of a subset $v \subseteq R_{X_i}$ is proportional to $\mu_p^{R_{X_i}}(v) = p^{|v|}(1-p)^{|R_{X_i} \setminus v|}$ as in Definition 2.4. All blocks in the same layer have the same total weight and the total weight of each layer is $\frac{1}{l}$. In other words, the weight of a vertex $v \in V[x]$ where $x \in X_i$ is given by

$$\frac{1}{l|X_i|} \mu_p^{R_{X_i}}(v).$$

**Hyperedges.** We construct hyperedges between blocks $V[x]$ and $V[y]$ such that there exists a constraint $\pi_{x\to y}$. We connect a hyperedge between any $v_1,\ldots,v_{k-1} \in V[x]$ and $u \in V[y]$ whenever $\pi_{x\to y}(\bigcap_{i=1}^{k-1} v_i) \cap u = \phi$.

Let $IS(G)$ denote the fraction of vertices contained in the largest independent set of the hypergraph $G$.

**Lemma 4.2 (Completeness)** *If $\Phi$ is satisfiable then $IS(G) \geq p$.*

**Proof:** Let $A$ be a satisfying assignment for $\Phi$, i.e., $A$ maps each $i \in [l]$ and $x \in X_i$ to an assignment in $R_{X_i}$ such that all the constraints are satisfied. Let $\mathcal{I} \subseteq V$ contain in the block $V[x]$ all the vertices that contain the assignment $A(x)$. We claim that $\mathcal{I}$ is an independent set. Take any $v_1,\ldots,v_{k-1}$ in $\mathcal{I} \cap V[x]$ and a vertex $u$ in $\mathcal{I} \cap V[y]$. The vertices $v_1,\ldots,v_{k-1}$ intersect on $A(x)$ and therefore the projection of their intersection contains $\pi_{x\to y}(A(x)) = A(y)$. Since $u$ is in $\mathcal{I} \cap V[y]$ it must contain $A(y)$. The proof is completed by noting that inside each block, the weight of the set of all vertices that contain a specific assignment is $p$. ■

The following lemma completes the proof of our main result since $\frac{1-\varepsilon}{1-p} = (1-\varepsilon)(k-1-\varepsilon)$ can be arbitrarily close to $k-1$.

**Lemma 4.3 (Soundness)** *If $IS(G) \geq \varepsilon$ then $\Phi$ is satisfiable.*

**Proof:** Let $\mathcal{I}$ be an independent set of weight $\varepsilon$. We consider the set $X'$ of all variables $x$ for which the weight of $\mathcal{I} \cap V[x]$ in $V[x]$ is at least $\varepsilon/2$. A simple averaging argument shows that the weight of $\bigcup_{x\in X'} V[x]$ is at least $\frac{\varepsilon}{2}$. Another averaging argument shows that in at least $\frac{\varepsilon}{4}l = \frac{8}{\varepsilon}$ layers, $X'$ contains at least $\frac{\varepsilon}{4}$ fraction of the variables. Using the weak-density property of the PCP (see Definition 3.2), we conclude that there exist two layers $X_i$ and $X_j$ such that $\frac{\varepsilon^2}{64}$ fraction of the constraints between them are constraints between variables in $X'$. Let us denote by $X$ the variables in $X_i \cap X'$ and by $Y$ the variables in $X_j \cap X'$.

For any variable $x \in X$, consider the vertices in $\mathcal{I} \cap V[x]$. According to Lemma 2.5 there exists a $t = t(\frac{\varepsilon}{2}, k-1, p)$ and $k-1$ vertices in $\mathcal{I} \cap V[x]$ that intersect in less than $t$ assignments. We denote these vertices by $v_{x,1},\ldots,v_{x,k-1}$ and their intersection by $B(x)$.

In the following we define an assignment to the variables in $X$ and $Y$ such that many of the constraints between them are satisfied. Then Theorem 3.3 would imply that $\Phi$ must be satisfiable (provided $R$ is chosen large enough). For a variable $x \in X$ we choose a random assignment from the set $B(x)$. For a variable $y \in Y$ we choose the assignment

$$A(y) = \text{maxvar}_{a\in R_Y} |\{x \in X \mid a \in \pi_{x\to y}(B(x))\}|,$$

i.e., the assignment that is contained in the largest number of projections of $B(x)$.

Before continuing, we need the following simple claim:

**Claim 4.4** *Let $A_1,\ldots,A_n$ be a collection of $n$ sets of size at most $m$ such that no element is contained in more than $k$ sets. Then, there are at least $\frac{n}{1+(k-1)m} \geq \frac{n}{km}$ disjoint sets in this collection.*

**Proof:** We prove by induction on $n$ that there are at least $\frac{n}{1+(k-1)m}$ disjoint sets in the collection. The claim holds trivially for $n \leq 1+(k-1)m$. Otherwise, consider all the sets that intersect $A_1$. Since no element is contained in more than $k$ sets, the number of such sets (including $A_1$) is at most $1+(k-1)m$. Removing these sets we get, by using the induction hypothesis, a collection that contains $\frac{n-1-(k-1)m}{1+(k-1)m} = \frac{n}{1+(k-1)m} - 1$ disjoint sets. We conclude the induction step by adding $A_1$ to the disjoint sets. ■

Consider a variable $y \in Y$ and a variable $x$ such that the constraint $\pi_{x \to y}$ exists. There are no hyperedges of the form $(v_{x,1}, \ldots, v_{x,k-1}, u)$ for any vertex $u \in \mathcal{I} \cap V[y]$. Therefore, every vertex $u \in \mathcal{I} \cap V[y]$ must intersect $\pi_{x \to y}(B(x))$. Now consider the family of projections $\pi_{x \to y}(B(x))$ for all the variables $x$ such that the constraint $\pi_{x \to y}$ exists. Let $q$ denote the maximum number of disjoint sets inside this family. Note that every disjoint set reduces the weight of the vertices in $\mathcal{I} \cap V[y]$ by a factor of $1 - (1-p)^t$. Because the weight of $\mathcal{I} \cap V[y]$ is at least $\frac{\varepsilon}{4}$, we obtain that $q$ is at most $\log(\frac{\varepsilon}{4}) / \log(1 - (1-p)^t)$. Claim 4.4 implies that there exists an assignment for $y$ that is contained in at least a fraction $\frac{1}{t \log(\frac{\varepsilon}{4}) / \log(1-(1-p)^t)}$ of the projections $\pi_{x \to y}(B(x))$. Therefore, the expected fraction of constraints satisfied between $X$ and $Y$ is at least $\frac{1}{t^2 \log(\frac{\varepsilon}{4}) / \log(1-(1-p)^t)}$ which is a constant that does not depend on $R$. We complete the proof by choosing the range $R$ of the PCP large enough and applying Theorem 3.3. This completes the soundness proof. ■

■

# 5 Acknowledgements

We would like to thank Noga Alon for his help with $s$-wise $t$-intersecting families.

## References


[1] S. Arora, C. Lund, R. Motwani, M. Sudan and M. Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3) : 501-555, 1998.

[2] S. Arora and S. Safra. Probabilistic checking of proofs : A new characterization of NP. *Journal of the ACM*, 45(1) : 70-122, 1998.

[3] M. Bellare, O. Goldreich and M. Sudan. Free bits, PCPs and non-approximability. *SIAM Journal on Computing*, 27(3):804-915, June 1998.

[4] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.

[5] I. Dinur, V. Guruswami and S. Khot. Vertex cover on $k$-uniform hypergraphs is hard to approximate within factor $(k - 3 - \varepsilon)$. *Electronic Colloquium on Computational Complexity, Technical Report TR02-027*, 2002.

[6] I. Dinur, O. Regev and C. Smyth. The hardness of 3-uniform hypergraph coloring. *In Proc. of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, November 2002, to appear.

[7] I. Dinur and S. Safra. The importance of being biased. *Proceedings of the 34th Annual ACM Symposium on Theory of Computing*, pages 33-42, May 2002.

[8] U. Feige. A threshold of ln n for approximating set cover. *JACM: Journal of the ACM*, 45, 1998.

[9] U. Feige, S. Goldwasser, L. Lovász, S. Safra and M. Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM*, 43(2):268-292, March 1996.

[10] O. Goldreich. Using the FGLSS-reduction to prove inapproximability results for minimum vertex cover in hypergraphs. *ECCC Technical Report TR01-102*, December 2001.

[11] V. Guruswami, J. Håstad and M. Sudan. Hardness of approximate hypergraph coloring. *Proceedings of the 41st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, pages 149-158, November 2000.

[12] E. Halperin. Improved approximation algorithms for the vertex cover problem in graphs and hypergraphs. *Proceedings of the 11th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 329-337, January 2000.

[13] R. L. Graham, M. Grötschel, and L. Lovász, editors. *Handbook of combinatorics. Vol. 1, 2.* Elsevier Science B.V., Amsterdam, 1995.

[14] J. Håstad. Clique is hard to approximate within $n^{1-\varepsilon}$. *Acta Mathematica*, 182(1):105-142, 1999.

[15] J. Håstad. Some optimal inapproximability results. *Journal of the ACM*, 48(4):798-859, July 2001.

[16] J. Holmerin. Vertex cover on 4-uniform hypergraphs is hard to approximate within $2 - \varepsilon$. *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 544-552, May 2002.

[17] J. Holmerin. Improved inapproximability results for vertex cover on k-uniform hypergraphs. *Proc. of the 29th International Colloquium on Automata, Languages and Programming (ICALP)*, pages 1005-1016, July 2002.

[18] D. S. Johnson. Approximation algorithms for combinatorial problems. *Journal of Computer and System Sciences*, 9:256–278, 1974.

[19] S. Khot. Hardness results for coloring 3-colorable 3-uniform hypergraphs. *In Proc. of the 43rd Annual IEEE Symposium on Foundations of Computer Science*, November 2002, to appear.

[20] L. Lovász. On the ratio of optimal integral and fractional covers. *Discrete Mathematics*, 13:383–390, 1975.

[21] R. Raz. A parallel repetition theorem. *SIAM J. of Computing*, vol 27(3), 763-803, 1998.

[22] A. Samorodnitsky and L. Trevisan. A PCP characterization of NP with optimal amortized query complexity. *In Proc. of the 32nd Annual ACM Symposium on Theory of Computing*, pages 191-199, 2000.

[23] L. Trevisan. Non-approximability results for optimization problems on bounded degree instances. *In Proc. of the Annual ACM 33rd Symposium on Theory of Computing*, pages 453-461, July 2001.

[24] V. V. Vazirani. *Approximation Algorithms.* Springer, 2001.