On the Approximability of Partial VC Dimension

Cristina Bazgan[1,⋆], Florent Foucaud[2], and Florian Sikora[1]

[1] Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, PARIS, FRANCE
{cristina.bazgan,florian.sikora}@dauphine.fr
[2] Université Blaise Pascal - CNRS UMR 6158 - LIMOS, Clermont-Ferrand, FRANCE
florent.foucaud@gmail.com

Abstract. We introduce the problem Partial VC Dimension that asks, given a hypergraph $H = (X, E)$ and integers k and ℓ, whether one can select a set $C \subseteq X$ of k vertices of H such that the set $\{e \cap C, e \in E\}$ of distinct hyperedge-intersections with C has size at least ℓ. The sets $e \cap C$ define equivalence classes over E. Partial VC Dimension is a generalization of VC Dimension, which corresponds to the case $\ell = 2^k$, and of Distinguishing Transversal, which corresponds to the case $\ell = |E|$ (the latter is also known as Test Cover in the dual hypergraph). We also introduce the associated fixed-cardinality maximization problem Max Partial VC Dimension that aims at maximizing the number of equivalence classes induced by a solution set of k vertices. We study the approximation complexity of Max Partial VC Dimension on general hypergraphs and on more restricted instances, in particular, neighborhood hypergraphs of graphs.

1 Introduction

We study identification problems in discrete structures. Consider a hypergraph (or set system) $H = (X, E)$, where X is the vertex set and E is a collection of hyperedges, that is, subsets of X. Given a subset $C \subseteq X$ of vertices, we say that two hyperedges of E are *distinguished* (or *separated*) by C if some element in C belongs to exactly one of the two hyperedges. In this setting, one can tell apart the two distinguished hyperedges simply by comparing their intersections with C. Following this viewpoint, one may say that two hyperedges are related if they have the same intersection with C. This is clearly an equivalence relation, and one may determine the collection of equivalence classes induced by C: each such class corresponds to its own subset of C. Any two hyperedges belonging to distinct equivalence classes are then distinguished by C. We call these classes *neighborhood equivalence classes*. In general, one naturally seeks to distinguish as many pairs of hyperedges as possible, using a small set C.

It is a well-studied setting to ask for a maximum-size set C such that C induces all possible $2^{|C|}$ equivalence classes. In this case, C is said to be *shattered*.

[⋆] Institut Universitaire de France

The maximum size of a shattered set in a hypergraph H is called its *Vapnis-Červonenkis dimension* (VC dimension for short). This notion, introduced by Vapnis and Červonenkis [42] arose in the context of statistical learning theory as a measure of the structural complexity of the data. It has since been widely used in discrete mathematics; see the references in the thesis [9] for more references. We have the following associated decision problem.

VC DIMENSION
Input: A hypergraph $H = (X, E)$, and an integer k.
Question: Is there a shattered set $C \subseteq X$ of size at least k in H?

The complexity of VC DIMENSION was studied in e.g. [17,21,36]; it is a complete problem for the complexity class LOGNP defined in [36] (it is therefore a good candidate for an NP-intermediate problem). VC DIMENSION remains LOGNP-complete for *neighborhood hypergraphs of graphs* [31] (the *neighborhood hypergraph* of G has $V(G)$ as its vertex set, and the set of closed neighborhoods of vertices of G as its hyperedge set).

In another setting, one wishes to distinguish *all* pairs of hyperedges (in other words, each equivalence class must have size 1) while minimizing the size of the solution set C. Following [27], we call the associated decision problem, DISTINGUISHING TRANSVERSAL.

DISTINGUISHING TRANSVERSAL
Input: A hypergraph $H = (X, E)$, and an integer k.
Question: Is there a set $C \subseteq X$ of size at most k that induces $|E|$ distinct equivalence classses?

There exists a rich literature about DISTINGUISHING TRANSVERSAL. It was studied under different names, such as TEST SET in Garey and Johnson's book [26, SP6]; other names include TEST COVER [18,19,20], DISCRIMINATING CODE [16] or SEPARATING SYSTEM [7,39].[3] A celebrated theorem of Bondy [8] also implicitely studies this notion. A version of DISTINGUISHING TRANSVERSAL called IDENTIFYING CODE was defined for graphs instead of hypergraphs [24,28]. Similarly as for the well-known relation between the classic graph problem DOMINATING SET and the hypergraph problem HITTING SET, it is easy to check that an identifying code in graph G is the same as a distinguishing transversal in the neighborhood hypergraph of G.

The goal of this paper is to introduce and study the problem PARTIAL VC DIMENSION, that generalizes both DISTINGUISHING TRANSVERSAL and VC DIMENSION, and defined as follows.

[3] Technically speaking, in TEST SET, TEST COVER and SEPARATING SYSTEM, the goal is to distinguish the *vertices* of a hypergraph using a set C of *hyperedges*, and in DISCRIMINATING CODE the input is presented as a bipartite graph. Nevertheless, these formulations are equivalent to DISTINGUISHING TRANSVERSAL by considering either the dual hypergraph of the input hypergraph $H = (X, E)$ (with vertex set E and hyperedge set X, and hyperedge x contains vertex e in the dual if hyperedge e contains vertex x in H), or the bipartite incidence graph (defined over vertex set $X \cup E$, and where x and e are adjacent if they were incident in H).

Partial VC Dimension
Input: A hypergraph $H = (X, E)$, and two integers k and ℓ.
Question: Is there a set $C \subseteq X$ of size k that induces at least ℓ distinct equivalence classses?

Partial VC Dimension belongs to the category of *partial* versions of common decision problems, in which, instead of satisfying the problem's constraint task for all elements (here, all 2^k equivalence classes), we ask whether we can satisfy a certain number, ℓ, of these constraints. See for example the papers [23,30] that study some partial versions of standard decision problems, such as Set Cover or Dominating Set.

When $\ell = |E|$, Partial VC Dimension is precisely the problem Distinguishing Transversal. When $\ell = 2^k$, we have the problem VC Dimension. Hence, Partial VC Dimension is NP-hard, even on many restricted classes. Indeed, Distinguishing Transversal is NP-hard [26], even on hypergraphs where each vertex belongs to at most two hyperedges [20], or on neighborhood hypergraphs of graphs that are either: unit disk graphs [34], planar bipartite subcubic [24], graphs that are interval and permutation [25], split graphs [24]. Min Distinguishing Transversal cannot be approximated within a factor of $o(\log n)$ on hypergraphs of order n [20], even on hypergraphs without 4-cycles [10], and on neighborhood hypergraphs of bipartite, co-bipartite or split graphs [24].

When $\ell = 2^k$, Partial VC Dimension is equivalent to VC Dimension and unlikely to be NP-hard (unless all problems in NP can be solved in quasi-polynomial time), since $|X| \leqslant 2^k$ and a simple brute-force algorithm has quasi-polynomial running time. Moreover, VC Dimension (and hence Partial VC Dimension) is W[1]-complete when parameterized by k [21].

Recently, the authors in [12] introduced the notion of (α, β)-*set systems*, that is, hypergraphs where, for any set S of vertices with $|S| \leqslant \alpha$, S induces at most β equivalence classes. Using this terminology, if a given hypergraph H is an (α, β)-set system, (H, k, ℓ) with $k = \alpha$ is a YES-instance of Partial VC Dimension if and only if $\ell \leqslant \beta$.

We will also study the approximation complexity of the following fixed-cardinality maximization problem associated to Partial VC Dimension.

Max Partial VC Dimension
Input: A hypergraph $H = (X, E)$, and an integer k.
Output: A set $C \subseteq X$ of size k that maximizes the number of equivalence classes induced by C.

Similar *fixed-cardinality* versions of classic optimization problems such as Set Cover, Dominating Set or Vertex Cover, derived from the "partial" counterparts of the corresponding decision problems, have gained some attention in the recent years, see for example [14,30,13].

Max Partial VC Dimension is clearly NP-hard since Partial VC Dimension is NP-complete; other than that, its approximation complexity is completely unknown since it cannot be directly related to the one of approximating

Min Distinguishing Transversal or Max VC Dimension (the minimization and maximization versions of Distinguishing Transversal and VC Dimension, respectively).

Our results. Our focus is on the approximation complexity of Max Partial VC Dimension. We give positive results in Section 3. We first provide polynomial-time approximation algorithms using the VC-dimension for the maximum degree or for the maximum edge-size of the input hypergraph. We apply these to obtain approximation ratios of the form n^{δ} (for $\delta < 1$ a constant) in certain special cases, as well as a better approximation ratio but with exponential time. For neighbourhood hypergraphs of planar graphs, Max Partial VC Dimension admits a PTAS (this is also shown for Min Distinguishing Transversal). In Section 4, we give hardness results. We show that any 2-approximation algorithm for Max Partial VC Dimension implies a 2-approximation algorithm for Max VC Dimension. Finally, we show that Max Partial VC Dimension is APX-hard, even for graphs of maximum degree at most 7.

2 Preliminaries

Twin-free hypergraphs. In a hypergraph H, we call two equal hyperedges *twin hyperedges*. Similarly, two vertices belonging to the same set of hyperedges are *twin vertices*.

Clearly, two twin hyperedges will always belong to the same neighborhood equivalence classes. Similarly, for any set T of mutually twin vertices, there is no advantage in selecting more than one of the vertices in T when building a solution set C.

Observation 1 *Let $H = (X, E)$ be a hypergraph and let $H' = (X', E')$ be the hypergraph obtained from H by deleting all but one of the hyperedges or vertices from each set of mutual twins. Then, for any set $C \subseteq X$, the equivalence classes induced by C in H are the same as those induced by $C \cap X'$ in H'.*

Therefore, since it is easy to detect twin hyperedges and vertices in an input hypergraph, in what follows, we will always restrict ourselves to hypergraphs without twins. We call such hypergraphs *twin-free*.

Degree conditions. In a hypergraph H, the *degree* of a vertex x is the number of hyperedges it belongs to. The *maximum degree* of H is the maximum value of the degree of a vertex of H; we denote it by $\Delta(H)$.

The next theorem gives an upper bound on the number of neighborhood equivalence classes that can be induced when the degrees are bounded.

Theorem 2 ([18,20,28]). *Let $H = (X, E)$ be a hypergraph with maximum degree Δ and let C be a subset of X of size k. Then, C cannot induce more than $\frac{k(\Delta+1)}{2} + 1$ neighborhood equivalence classes.*

The Sauer-Shelah lemma. The following theorem is known as the Sauer-Shelah Lemma [40,41] (it is also credited to Perles in [41] and a weaker form was stated by Vapnik and Červonenkis [42]). It is a fundamental tool in the study of the VC dimension.

Theorem 3 (Sauer-Shelah Lemma [40,41]). *Let $H = (X, E)$ be a hypergraph with strictly more than $\sum_{i=0}^{d-1} \binom{|X|}{i}$ distinct hyperedges. Then, S has VC-dimension at least d.*

Theorem 3 is known to be tight. Indeed, the system that consists of considering all subsets of $\{1, \ldots, n\}$ of cardinality at most $d-1$ has VC-dimension equal to $d-1$. Though the original proofs of Theorem 3 were non-constructive, Ajtai [1] gave a constructive proof that yields a (randomized) polynomial-time algorithm, and an easier proof of this type can be found in Micciano [32].

The following direct corollary of Theorem 3 is observed for example in [10].

Corollary 4. *Let $S = (X, E)$ be a hypergraph with VC dimension at most d. Then, for any subset $X' \subseteq X$, there are at most $\sum_{i=0}^{d} \binom{|X'|}{i} \leqslant |X'|^d + 1$ equivalence classes induced by X'.*

Approximation. An algorithm for an optimization problem is a c-approximation algorithm if it returns a solution whose value is always at most a factor of c away from the optimum. The class APX contains all optimization problems that admit a polynomial-time c-approximation algorithm for some fixed constant c. A *polynomial-time approximation scheme* (PTAS for short) for an optimization problem is an algorithm that, given any fixed constant $\epsilon > 0$, returns in polynomial time (in terms of the instance and for fixed ϵ) a solution that is a factor of $1+\epsilon$ away from the optimum. An optimization problem is APX-*hard* if it admits no PTAS (unless P=NP).

Given an optimization problem P, an instance I of P, we denote by $opt_P(I)$ (or $opt(I)$ if there is no ambiguity) the value of an optimal solution for I.

Definition 5 (L-reduction [35]). *Let A and B be two optimization problems. Then A is said to be L-reducible to B if there are two constants $\alpha, \beta > 0$ and two polynomial time computable functions f, g such that: (i) f maps an instance I of A into an instance I' of B such that $opt_B(I') \leqslant \alpha \cdot opt_A(I)$, (ii) g maps each solution S' of I' into a solution S of I such that $||S| - opt_A(I)| \leqslant \beta \cdot ||S'| - opt_B(I')|$.*

L-reductions are useful in order to apply the following theorem.

Theorem 6 ([35]). *Let A and B be two optimization problems. If A is* APX-*hard and L-reducible to B, then B is* APX-*hard.*

3 Positive approximation results for Max Partial VC Dimension

We start with a greedy polynomial-time procedure that always returns (if it exists), a set $|X'|$ that induces at least $|X'|+1$ equivalence classes.

Lemma 7. *Let $H = (X, E)$ be a twin-free hypergraph and let $k \leqslant |X| - 1$ be an integer. One can construct, in time $O(k(|X| + |E|))$, a set $C \subseteq X$ of size k that produces at least $\min\{|E|, k+1\}$ neighborhood equivalence classes.*

Proof. We produce C in an inductive way. First, let $C_1 = \{x\}$ for an arbitrary vertex x of X for which there exists at least one hyperedge of E with $x \notin E$ (if such hyperedge does not exist, then all edges are twin edges; since H is twin-free, $|E| \leqslant 1$ and we are done). Then, for each i with $2 \leqslant i \leqslant k$, we build C_i from C_{i-1} as follows: select vertex x_i as a vertex in $X \setminus C_{i-1}$ such that $C_{i-1} \cup \{x\}$ maximizes the number of equivalence classes.

We claim that either we already have at least $|E|$ equivalence classes, or C_i induces at least one more equivalence class than C_{i-1}. Assume for a contradiction that we have strictly less than $|E|$ equivalence classes, but C_i has the same number of equivalence classes as C_{i-1}. Since we have strictly less than $|E|$ classes, there is an equivalence class consisting of at least two edges, say e_1 and e_2. But then, since H is twin-free, there is a vertex x that belongs to exactly one of e_1 and e_2. But $C_{i-1} \cup \{x\}$ would have strictly more equivalence classes than C_i, a contradiction since C_i was maximizing the number of equivalence classes.

Hence, setting $C = C_k$ finishes the proof. □

Proposition 8. Max Partial VC Dimension *is $\frac{\min\{2^k, |E|\}}{k+1}$-approximable in polynomial time. For hypergraphs with VC dimension at most d,* Max Partial VC Dimension *is k^{d-1}-approximable. For hypergraphs with maximum degree Δ,* Max Partial VC Dimension *is $\frac{\Delta+1}{2}$-approximable.*

Proof. By Lemma 7, we can always compute in polynomial time, a solution with at least $k+1$ neighborhood equivalence classes (if it exists; otherwise, we solve the problem exactly). Since there are at most $\min\{2^k, |E|\}$ possible classes, the first part of the statement follows. Similarly, by Corollary 4, if the hypergraph has VC dimension at most d, there are at most $k^d + 1$ equivalence classes, and $\frac{k^d+1}{k+1} \leqslant k^{d-1}$. Finally, if the maximum degree is at most Δ, by Theorem 2 there are at most $\frac{k(\Delta+1)+2}{2}$ possible classes (and when $\Delta \geqslant 1$, $\frac{k(\Delta+1)+2}{2(k+1)} \leqslant \frac{\Delta+1}{2}$). □

Corollary 9. *For hypergraphs of VC dimension at most d,* Max Partial VC Dimension *is $|E|^{(d-1)/d}$-approximable.*

Proof. By Proposition 8, we have a $\min\{k^{d-1}, |E|/k\}$)-approximation. If $k^{d-1} < |E|^{(d-1)/d}$ we are done. Otherwise, we have $k^{d-1} \geqslant |E|^{(d-1)/d}$ and hence $k \geqslant |E|^{1/d}$, which implies that $\frac{|E|}{k} \leqslant |E|^{(d-1)/d}$. □

For examples of concrete applications of Corollary 9, hypergraphs with no 4-cycles in its bipartite incidence graph[4] have VC-dimension at most 3 and hence we have an $|E|^{2/3}$-approximation for this class. Hypergraphs with maximum edge-size d also have VC dimension at most d. Other examples, arising from

[4] In the dual hypergraph, this corresponds to the property that each pair of hyperedges have at most one common element, see for example [2].

graphs, are neighborhood hypergraphs of: K_{d+1}-minor-free graphs (that have VC dimension at most d [11]); graphs of rankwidth at most r (VC dimension at most $2^{2^{O(r)}}$ [11]); interval graphs (VC dimension at most 2 [10]); permutation graphs (VC dimension at most 3 [10]); line graphs (VC dimension at most 3); unit disk graphs (VC dimension at most 3) [10]; C_4-free graphs (VC dimension at most 2); chordal bipartite graphs (VC dimension at most 3 [10]); undirected path graphs (VC dimension at most 3 [10]). Typical graph classes with unbounded VC dimension are bipartite graphs and their complements, or split graphs.

In the case of hypergraphs with no 4-cycles in its bipartite incidence graph (for which Max Partial VC Dimension has an $|E|^{2/3}$-approximation algorithm by Corollary 9), we can also relate Max Partial VC Dimension to Max Partial Double Hitting Set, defined as follows.

Max Partial Double Hitting Set
Input: A hypergraph $H = (X, E)$, an integer k.
Output: A subset $C \subseteq X$ of size k maximizing the number of hyperedges containing at least two elements of C.

Theorem 10. *Any α-approximation algorithm for* Max Partial Double Hitting Set *on hypergraphs without 4-cycles in its bipartite incidence graph can be used to obtain a 4α-approximation algorithm for* Max Partial VC Dimension *on hypergraphs without 4-cycles in its bipartite incidence graph.*

Proof. Let $H = (X, E)$ be a hypergraph without 4-cycles in its bipartite incidence graph, and let $C \subseteq X$ be a subset of vertices. Since H has no 4-cycles in its bipartite incidence graph, note that if some hyperedge contains two vertices of X, then no other hyperedge contains these two vertices. Therefore, the number of equivalence classes induced by C is equal to the number of hyperedges containing at least two elements of C, plus the number of equivalence classes corresponding to a single (or no) element of C. Therefore, the maximum number $opt(H)$ of equivalence classes for a set of size k is at most $opt_{2HS}(H) + k + 1$, where $opt_{2HS}(H)$ is the value of an optimal solution for Max Partial Double Hitting Set on H. Observing that $opt_{2HS}(H) \geqslant \frac{k}{2}$ (since one may always iteratively select pairs of vertices covering a same hyperedge to obtain a valid double hitting set of H), we get that $opt(H) \leqslant 3opt_{2HS}(H) + 1 \leqslant 4opt_{2HS}(H)$. Moreover, in polynomial time we can apply the approximation algorithm of Max Partial Double Hitting Set to H to obtain a set C inducing at least $\frac{opt_{2HS}(H)}{\alpha}$ neighborhood equivalence classes. Thus, C induces at least $\frac{opt(H)}{4\alpha}$ neighborhood equivalence classes. □

Unfortunately, the complexity of approximating Max Partial Double Hitting Set seems not to be well-known, even when restricted to hypergraphs with no 4-cycles in its bipartite incidence graph. In fact, the problem Max Densest Subgraph (which, given an input graph, consists of maximizing the number of edges of a subgraph of order k) is precisely Max Partial Double Hitting Set restricted to hypergraphs where each hyperedge has size at most 2 (that is, to graphs), that can be assumed to contain no 4-cycles in its bipartite incidence graph (a 4-cycle would imply the existence of two twin hyperedges). Although Max Densest Subgraph (and hence Max Partial Double Hitting Set for hypergraphs with no 4-cycles in its bipartite incidence graph) is only known to admit no PTAS [29], the best known approximation ratio for it is $O(|E|^{1/4})$ [5].[5] We deduce from this result, the following corollary of Theorem 10 for hypergraphs of hyperedge-size bounded by 2. This improves on the $O(|E|^{1/2})$-approximation algorithm given by Corollary 9 for this case.

Corollary 11. *Let α be the best approximation ratio in polynomial time for* Max Densest Subgraph. *Then,* Max Partial VC Dimension *can be 3α-approximated in polynomial time on hypergraphs with hyperedges of size at most 2. In particular, there is a polynomial-time $O(|E|^{1/4})$-approximation algorithm for this case.*

We will now apply the following result from [4].

Lemma 12 ([4]). *If an optimization problem is $r_1(k)$-approximable in fpt-time with respect to parameter k for* some *strictly increasing function r_1 depending solely on k, then it is also $r_2(n)$-approximable in fpt-time w.r.t. parameter k for* any *strictly increasing function r_2 depending solely on the instance size n.*

Using Proposition 8 showing that Max Partial VC Dimension is $\frac{2^k}{k+1}$-approximable and Lemma 12, we directly obtain the following.

Corollary 13. *For any strictly increasing function r,* Max Partial VC Dimension *parameterized by k is $r(n)$-approximable in FPT-time.*

In the following we establish polynomial time approximation schemes for Min Distinguishing Transversal and Max Partial VC Dimension on planar graphs using the layer decomposition technique introduced by Baker [3].

Given a planar embedding of an input graph, we call the vertices which are on the external face *level 1 vertices*. By induction, we define *level t vertices* as the set of vertices which are on the external face after removing the vertices of levels smaller than t [3]. A planar embedding is *t-level* if it has no vertices of level greater than t. If a planar graph is t-level, it has a t-outerplanar embedding.

Theorem 14. Max Partial VC Dimension *on neighborhood hypergraphs of planar graphs admits a PTAS.*

Proof. Let G be a planar graph with a t-level planar embedding for some integer t. We aim to achieve an approximation ratio of $1+\varepsilon$. Let $\lambda = \lceil \frac{1}{\varepsilon} \rceil - 1$.

Let G_i $(0 \leqslant i \leqslant \lambda)$ be the graph obtained from G by removing the vertices on levels $i \bmod (\lambda+1)$. Thus, graph G_i is the disjoint union of several subgraphs G_{ij} $(0 \leqslant j \leqslant p$ with $p = \lceil \frac{t+i}{\lambda+1} \rceil)$ where G_{i0} is induced by the vertices on

[5] Formally, it is stated in [5] as an $O(|V|^{1/4})$-approximation algorithm, but we may assume that the input graph is connected, and hence $|V| = O(|E|)$.

levels $0, \dots, i-1$ (note that G_{00} is empty) and G_{ij} with $j \geqslant 1$ is induced by the vertices on levels $(j-1)(\lambda+1)+i+1, \dots j(\lambda+1)+i-1$. In other words, each subgraph G_{ij} is the union of at most λ consecutive levels and is thus λ-outerplanar. Hence, G_i is also λ-outerplanar and it has treewidth at most $3\lambda-1$ [6]. Using Courcelle's theorem[6], for any integer t and any subgraph G_{ij}, we can efficiently determine an optimal set S^t_{ij} of t vertices of G_{ij} that maximizes the number of (nonempty) induced equivalence classes in G_{ij}. We then use dynamic programming to construct a solution for G_i. Denote by $S_i(q, y)$ a solution corresponding to the maximum feasible number of equivalence classes induced by a set of y vertices of G_i ($0 \leqslant y \leqslant k$) among the first q subgraphs $G_{i1}, \dots, G_{iq}$ ($1 \leqslant q \leqslant p$). We have $S_i(q, y) = \max_{0 \leqslant x \leqslant y}(S^x_{iq} + S_i(q-1, y-x))$. Let $S_i = S_i(p, k)$.

Among $S_0, \dots, S_\lambda$, we choose the best solution, that we denote by S. We now prove that S is an $(1+\varepsilon)$-approximation of the optimal value $opt(G)$ for Max Partial VC Dimension on G. Let S_{opt} be an optimal solution of G. Then, there is at least one integer r such that at most $1/(\lambda+1)$ of the equivalent classes induced by S_{opt} in G are lost when we remove vertices on the levels congruent to r mod $(\lambda+1)$.

Thus, $val(S) \geqslant val(S_r) \geqslant opt(G) - \frac{opt(G)}{\lambda+1} = \frac{\lambda}{\lambda+1}opt(G) \geqslant (1-\frac{1}{\varepsilon})opt(G)$, which completes the proof.

The overall running time of the algorithm is λ times the running time for graphs of treewidth at most $3\lambda-1$, that is, $O(\lambda n)$. □

As a side result, using the same technique, we provide the following theorem about Min Distinguishing Transversal, which is an improvement over the 7-approximation algorithm that follows from [38] (in which it is proved that any YES-instance satisfies $\ell \leqslant 7k$) and solves an open problem from [24]. Due to space constraints, its proof is omitted.

Theorem 15. Min Distinguishing Transversal *on neighborhood hypergraphs of planar graphs (equivalently,* Min Identifying Code *on planar graphs) admits a PTAS.*

4 Hardness of approximation results for Max Partial VC Dimension

We define Max VC Dimension as the maximization version of VC Dimension.

Max VC Dimension
Input: A hypergraph $H = (X, E)$.
Output: A maximum-size shattered subset $C \subseteq X$ of vertices.

[6] We can indeed encode the decision version of our problem in MSOL as follows: $\exists x_1, \dots, x_k, y_1, \dots, y_l, s^1_1, s^2_1, \dots, s^\ell_{\ell-1}, s^j_i = \bigvee_{q=1}^k x_q, \bigvee_{i,j=0}^{\binom{\ell}{2}} (s^j_i \in y_i \wedge s^j_i \notin y_j) \vee (s^j_i \notin y_i \wedge s^j_i \in y_j)$.

Not much is known about the complexity of Max VC Dimension: it is trivially $\log_2 |E|$-approximable by returning a single vertex; a lower bound on the running time of a potential PTAS has been proved [17]. It is mentioned as an outstanding open problem in [15]. In the following we establish a connection between the approximability of Max VC Dimension and Max Partial VC Dimension.

Theorem 16. *Any* 2*-approximation algorithm for* Max Partial VC Dimension *can be transformed into a randomized* 2*-approximation algorithm for* Max VC Dimension *with polynomial overhead in the running time.*

Proof. Let H be a hypergraph on n vertices that is an instance for Max VC Dimension, and suppose we have a c-approximation algorithm $\mathscr{A}$ for Max Partial VC Dimension.

We run $\mathscr{A}$ with $k = 1, \ldots, \log_2 |X|$, and let k_0 be the largest value of k such that the algorithm outputs a solution with at least $\frac{2^k}{c}$ neighborhood equivalence classes. Since $\mathscr{A}$ is a c-approximation algorithm, we know that the optimum for Max Partial VC Dimension for any $k > k_0$ is strictly less than 2^k. This implies that the VC-dimension of S is at most k_0.

Now, let X be the solution set of size k_0 computed by $\mathscr{A}$, and let H_X be the sub-hypergraph of H induced by X. By our assumption, this hypergraph has at least $\frac{2^{k_0}}{c}$ distinct edges. We can now apply the Sauer-Shelah Lemma (Theorem 3).

We have $c = 2$, and we apply the lemma with $|X| = k_0$ and $d = \frac{k_0}{2} + 1$; it follows that the VC dimension of H_X (and hence, of H) is at least $\frac{k_0}{2} + 1$. By the constructive proof of Theorem 3, a shattered set Y of this size can be computed in (randomized) polynomial time [1,32]. Set Y is a 2-approximation, since we saw in the previous paragraph that the VC dimension of H is at most k_0. □

We note that the previous proof does not seem to apply for any other constant than 2, because the Sauer-Shelah Lemma would not apply. Though the approximation complexity of Max VC Dimension is not known, our result shows that Max Partial VC Dimension is at least as hard to approximate.

Before proving our next result, we first need an intermediate result for Max Partial Vertex Cover (also known as Max k-Vertex Cover [14]), which is defined as follows.

Max Partial Vertex Cover
Input: A graph $G = (V, E)$, an integer k.
Output: A subset $S \subseteq V$ of size k covering the maximum number of edges.

Proposition 17 ([37]). Max Partial Vertex Cover *is* APX*-hard, even for cubic graphs.*

Theorem 18. Max Partial VC Dimension *is* APX*-hard, even for graphs of maximum degree* 7*.*

Proof. We will give an L-reduction from Max Partial Vertex Cover (which is APX-hard, by Proposition 17) to Max Partial VC Dimension. The result will then follow from Theorem 6. Given an instance $I = (G, k)$ of Max Partial Vertex Cover with $G = (V, E)$ a cubic graph, we construct an instance $I' = (G', k')$ of Max Partial VC Dimension with $G' = (V', E')$ of maximum degree 7 in the following way. For each vertex $v \in V$, we create a gadget P_v with twelve vertices where four among these twelve vertices are special: they form the set $F_v = \{f_v^1, f_v^2, f_v^3, f_v^4\}$. The other vertices are adjacent to the subsets $\{f_v^4\}$, $\{f_v^2, f_v^3\}$, $\{f_v^1, f_v^3\}$, $\{f_v^2, f_v^4\}$, $\{f_v^1, f_v^4\}$, $\{f_v^1, f_v^3, f_v^4\}$, $\{f_v^1, f_v^2, f_v^4\}$,$\{f_v^1, f_v^2, f_v^3, f_v^4\}$, respectively. We also add edges between f_v^1 and f_v^2, between f_v^2 and f_v^3 and between f_v^3 and f_v^4. Since G is cubic, for each vertex v of G, there are three edges e_1, e_2 and e_3 incident with v. For each edge e_i $(1 \leqslant i \leqslant 3)$, the endpoint v is replaced by f_v^i. Moreover, each of these original edges of G is replaced in G' by two edges by subdividing it once (see Figure 1 for an illustration). We call the vertices resulting from the subdivision process, *edge-vertices*. Finally, we set $k' = 4k$.

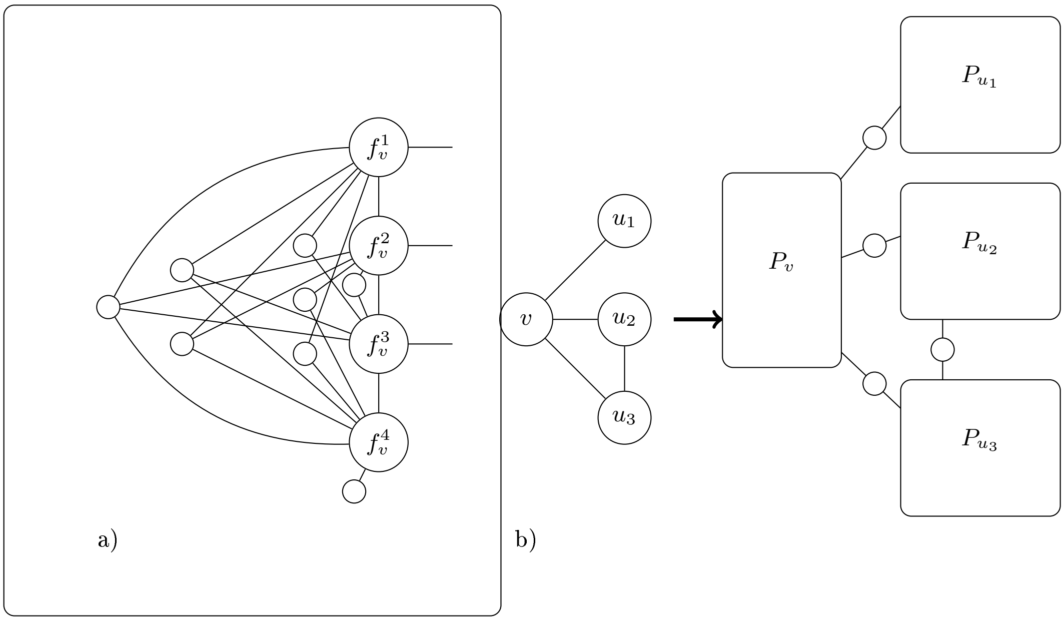

Fig. 1. a) Vertex-gadget P_v and b) illustration of the reduction.

From any optimal solution S with $|S| = k$ covering $opt(I)$ edges of G, we construct a set $C = \{f_v^j : 1 \leqslant j \leqslant 4, v \in S\}$ of size $4k$. By construction, C induces 12 equivalence classes in each vertex gadget. Moreover, for each covered edge $e = xy$ in G, the corresponding edge-vertex v_e in G' forms a class of size 1 (which corresponds to one or two neighbor vertices f_x^i and f_y^j of v_e in C).

Finally, all vertices in G' corresponding to edges not covered by S in G, as well as all vertices in vertex gadgets corresponding to vertices not in S, belong to the same equivalence class (corresponding to the empty set). Thus, C induces in G' $12k + opt(I) + 1$ equivalence classes, and hence we have

$$opt(I') \geqslant 12k + opt(I) + 1. \tag{1}$$

Conversely, given a solution C' of I' with $|C'| = 4k$, we transform it into a solution for I as follows. First, we show that C' can be transformed into another solution C'' such that (1) C'' only contains vertices of the form f_v^i, (2) each vertex-gadget contains either zero or four vertices of C'', and (3) C'' does not induce less equivalence classes than C'. To prove this, we proceed step by step by locally altering C' whenever (1) and (2) are not satisfied, while ensuring (3).

Suppose first that some vertex-gadget P_v of G' contains at least four vertices of C'. Then, the number of equivalence classes involving some vertex of $V(P_v) \cap C'$ is at most twelve within P_v (since there are only twelve vertices in P_v), and at most three outside P_v (since there are only three vertices not in P_v adjacent to vertices in P_v). Therefore, we can replace $V(P_v) \cap C'$ by the four special vertices of the set F_v in P_v; this choice also induces twelve equivalence classes within P_v, and does not decrease the number of induced classes.

Next, we show that it is always best to select the four special vertices of F_v from some vertex-gadget (rather than having several vertex-gadgets containing less than four solution vertices each). To the contrary, assume that there are two vertex-gadgets P_u and P_v containing respectively a and b vertices of C', where $1 \leqslant b \leqslant a \leqslant 3$. Then, we remove an arbitrary vertex from $C' \cap V(P_v)$; moreover we replace $C' \cap V(P_u)$ with the subset $\{f_u^i, 1 \leqslant i \leqslant a+1\}$, and similarly we replace $C' \cap V(P_v)$ with the subset $\{f_v^i, 1 \leqslant i \leqslant b-1\}$. Before this alteration, the solution vertices within $V(P_u) \cup V(P_v)$ could contribute to at most $2^a + 2^b - 2$ equivalence classes. After the modification, one can check that this quantity is at least $2^{a+1} + 2^{b-1} - 2$ classes. Observing that $2^{a+1} + 2^{b-1} - 2 \geqslant 2^a + 2^b - 2$ since $2^a - 2^{b-1} \geqslant 0$ yields our claim. Hence, by this argument, we conclude that all vertex-gadgets (except possibly at most one) contain either zero or four vertices from the solution set C'.

Suppose that there exists one vertex-gadget P_v with i solution vertices, $1 \leqslant i \leqslant 3$. We show that we may add $4-i$ solution vertices to it so that $C' \cap V(P_v) = F_v$. Consider the set of edge-vertices belonging to C'. Since we had $|C'| = 4k$ and all but one vertex-gadget contain exactly four solution vertices, there are at least $4-i$ edge-vertices in the current solution set. Then, we remove an arbitrary set of $4-i$ edge-vertices from C' and instead, we replace the set $V(P_v) \cap C'$ by the set F_v of special vertices of P_v. We now claim that this does not decrease the number of classes induced by C'. Indeed, any edge-vertex, since it has degree 2, may contribute to at most three equivalence classes, and the i solution vertices in P_v can contribute to at most 2^i classes. Summing up, in the old solution set, these four vertices contribute to at most $3(4-i) + 2^i$ classes, which is less than 12 since $1 \leqslant i \leqslant 3$. In the new solution, these four vertices contribute to at least 12 classes, which proves our above claim.

We now know that there are $4i$ edge-vertices in C', for some $i \leqslant k$. All other solution vertices are special vertices in some vertex-gadgets. By similar arguments as in the previous paragraph, we may select any four of them and replace them with some set F_v of special vertices of some vertex-gadget P_v. Before this modification, these four solution vertices may have contributed to at most $3 \cdot 4 = 12$ classes, while the new four solution vertices now contribute to at least 12 classes.

Applying the above arguments, we have proved the existence of the required set C'' that satisfies conditions (1)–(3).

Therefore, we may now assume that the solution C'' contains no edge-vertices, and for each vertex-gadget P_v, $C'' \cap V(P_v) \in \{\emptyset, F_v\}$. We define as solution S for I the set of vertices v of G for which P_v contains four vertices of C''. Then, $val(S) = val(C') - 12k - 1$. Considering an optimal solution C' for I', we have $opt(I) \geqslant opt(I') - 12k - 1$. Using (1), we conclude that $opt(I') = opt(I) + 12k + 1 \leqslant opt(I) + 24opt(I) + 1$ since $k \leqslant 2opt(I)$ and thus $opt(I') \leqslant 26opt(I)$.

Moreover, we have $opt(I) - val(S) = opt(I') - 12k - 1 - (val(C') - 12k - 1) = opt(I') - val(C')$.

Thus, our reduction is an L-reduction with $\alpha = 26$ and $\beta = 1$. □

Proposition 8 and Theorem 18 give the following corollary:

Corollary 19. Max Partial VC Dimension *is* APX*-complete for bounded degree graphs.*

5 Conclusion

In this paper, we defined and studied generalization of Distinguishing Transversal and VC Dimension. The probably most intriguing open question seems to be the approximation complexity of Max Partial VC Dimension. In particular, does the problem admit a constant-factor approximation algorithm? As a first step, one could determine whether such an approximation algorithm exists in superpolynomial time, or on special subclasses such as neighbourood hypergraphs of specific graphs. We have seen that there exist polynomial-time approximation algorithms with a sublinear ratio for special cases; does one exist in the general case?

References

1. M. Ajtai. The shortest vector problem in L2 is NP-hard for randomized reductions. Proc. of the 30th annual ACM Symposium on Theory of Computing (STOC'98):10–19, 1998.
2. V. S. Anil Kumar, S. Arya and H. Ramesh. Hardness of Set Cover with intersection 1. Proc. of the 27th International Colloquium on Automata, Languages and Programming (ICALP'00), LNCS 1853:624–635, 2000.
3. B. S. Baker. Approximation algorithms for NP-complete problems on planar graphs. *Journal of the ACM* 41(1):153–180, 1994.

4. C. Bazgan, M. Chopin, A. Nichterlein and F. Sikora. Parameterized approximability of maximizing the spread of influence in networks. *Journal of Discrete Algorithms* 27:54–65, 2014.
5. A. Bhaskara, M. Charikar, E. Chlamtac, U. Feige and A. Vijayaraghavan. Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph. Proc. of the 42nd annual ACM Symposium on Theory of Computing (STOC'10): 201–210, 2010.
6. H. Bodlaender. A partial k-arboretum of graphs with bounded treewidth. *Theoretical Computer Science* 209(12):1–45, 1998.
7. B. Bollobás and A. D. Scott. On separating systems. *European Journal of Combinatorics* 28:1068–1071, 2007.
8. J. A. Bondy. Induced subsets. *Journal of Combinatorial Theory, Series B* 12(2):201–202, 1972.
9. N. Bousquet. *Hitting sets: VC-dimension and Multicut.* PhD Thesis, Université Montepellier II, France, 2013. Available online at `http://tel.archives-ouvertes.fr/tel-01012106/`
10. N. Bousquet, A. Lagoutte, Z. Li, A. Parreau and S. Thomassé. Identifying codes in hereditary classes of graphs and VC-dimension. *SIAM Journal on Discrete Mathematics* 29(4):2047–2064, 2015.
11. N. Bousquet and S. Thomassé. VC-dimension and Erdős-Pósa property of graphs. *Discrete Mathematics* 338(12):2302–2317, 2015.
12. K. Bringmann, L. Kozma, S. Moran and N. S. Narayanaswamy. Hitting Set in hypergraphs of low VC-dimension. Manuscript, 2015. `http://arxiv.org/abs/1512.00481`
13. M. Bruglieri, M. Ehrgott, H. W. Hamacher and F. Maffioli. An annotated bibliography of combinatorial optimization problems with fixed cardinality constraints. *Discrete Applied Mathematics* 154(9):1344–1357, 2006.
14. L. Cai. Parameterized complexity of cardinality constrained optimization problems. *The Computer Journal* 51(1):102–121, 2007.
15. L. Cai, D. Juedes and I. Kanj. The inapproximability of non-NP-hard optimization problems. *Theoretical Computer Science* 289(1):553–571, 2002.
16. E. Charbit, I. Charon, G. Cohen, O. Hudry and A. Lobstein. Discriminating codes in bipartite graphs: bounds, extremal cardinalities, complexity. *Advances in Mathematics of Communications* 2(4):403–420, 2008.
17. J. Chen, X. Huang, I. A. Kanj and G. Xia. On the computational hardness based on linear FPT-reductions. *Journal of Combinatorial Optimization* 11(2):231–247, 2006.
18. R. Crowston, G. Gutin, M. Jones, G. Muciaccia and A. Yeo. Parameterizations of test cover with bounded test sizes. *Algorithmica* 74(1):367–384, 2016.
19. R. Crowston, G. Gutin, M. Jones, S. Saurabh and A. Yeo. Parameterized study of the test cover problem. Proc. of the 37th International Symposyium on Mathematical Foundations of Computer Science (MFCS'12), LNCS, 7464:283–295, 2012.
20. K. M. J. De Bontridder, B. V. Halldórsson, M. M. Halldórsson, C. A. J. Hurkens, J. K. Lenstra, R. Ravi and L. Stougie. Approximation algorithms for the test cover problem. *Mathematical Programming Series B* 98:477–491, 2003.
21. R. G. Downey, P. A. Evans and M. R. Fellows. Parameterized learning complexity. Proc. of the 6th annual conference on Computational learning theory (COLT'93), 51–57, 1993.
22. R. G. Downey and M. R. Fellows. *Fundamentals of Parameterized Complexity.* Springer, 2013.

23. F. V. Fomin, D. Lokshtanov, V. Raman and S. Saurabh. Subexponential algorithms for partial cover problems. *Information Processing Letters* 111(16):814–818, 2011.
24. F. Foucaud. Decision and approximation complexity for identifying codes and locating-dominating sets in restricted graph classes. *Journal of Discrete Algorithms* 31:48–68, 2015.
25. F. Foucaud, G. B. Mertzios, R. Naserasr, A. Parreau and P. Valicov. Algorithms and complexity for metric dimension and location-domination on interval and permutation graphs. Proc. of the 41st International Workshop on Graph-Theoretic Concepts in Computer Science (WG'15), LNCS 9224, to appear.
26. M. R. Garey and D. S. Johnson. *Computers and intractability: a guide to the theory of NP-completeness*, W. H. Freeman, 1979.
27. M. A. Henning and A. Yeo. Distinguishing-transversal in hypergraphs and identifying open codes in cubic graphs. *Graphs and Combinatorics* 30(4):909–932, 2014.
28. M. G. Karpovsky, K. Chakrabarty and L. B. Levitin. On a new class of codes for identifying vertices in graphs. *IEEE Transactions on Information Theory* 44:599–611, 1998.
29. S. Khot. Ruling Out PTAS for Graph Min-Bisection, Dense k-Subgraph, and Bipartite Clique. *SIAM J. Comput.* 36(4): 1025-1071, 2006.
30. J. Kneis, D. Mölle and P. Rossmanith. Partial vs. complete domination: t-dominating set. Proc. of the 33nd Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM'07), 367–376, 2007
31. E. Kranakis, D. Krizanc, B. Ruf, J. Urrutia and G. J. Woeginger. The VC-dimension of set systems defined by graphs. *Discrete Applied Mathematics* 77(3):237–257, 1997.
32. D. Micciancio. The shortest vector in a lattice is hard to approximate to within some constant. *SIAM Journal on Computing* 30(6):2008–2035, 2001.
33. J. Moncel. *Codes Identifiants dans les Graphes.* PhD Thesis, Université Joseph-Fourier - Grenoble I, France, June 2005. Available online at http://tel.archives-ouvertes.fr/tel-00010293.
34. T. Müller and J.-S. Sereni. Identifying and locating-dominating codes in (random) geometric networks. *Combinatorics, Probability and Computing* 18(6):925–952, 2009.
35. C. H. Papadimitriou and M. Yannakakis. Optimization, approximation, and complexity classes. *Journal of Computer and System Sciences* 43(3):425–440, 1991.
36. C. H. Papadimitriou and M. Yannakakis. On limited nondeterminism and the complexity of the V-C dimension. *Journal of Computer and System Sciences* 53(2):161–170, 1996.
37. E. Petrank. The hardness of approximation: gap location. *Computational Complexity* 4:133–157, 1994.
38. P. J. Slater and D. F. Rall. On location-domination numbers for certain classes of graphs. *Congressus Numerantium* 45:97–106, 1984.
39. A. Rényi. On random generating elements of a finite Boolean algebra. *Acta Scientiarum Mathematicarum Szeged* 22:75–81, 1961.
40. N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A* 13:145–147, 1972.
41. S. Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics* 41:247–261, 1972.
42. V. N. Vapnik and A. J. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Akademija Nauk SSSR* 16:264–279, 1971.