From Data Completion to Problems on
Hypercubes: A Parameterized Analysis of the
Independent Set Problem

Eduard Eiben
Department of Computer Science, Royal Holloway, University of London, Egham, UK

Robert Ganian
Algorithms and Complexity Group, TU Wien, Austria

Iyad Kanj
School of Computing, DePaul University, Chicago, IL, USA

Sebastian Ordyniak
School of Computing, University of Leeds, UK

Stefan Szeider
Algorithms and Complexity Group, TU Wien, Austria

Abstract
Several works have recently investigated the parameterized complexity of data completion problems,
motivated by their applications in machine learning, and clustering in particular. Interestingly, these
problems can be equivalently formulated as classical graph problems on induced subgraphs of powers
of partially-defined hypercubes.

In this paper, we follow up on this recent direction by investigating the Independent Set problem
on this graph class, which has been studied in the data science setting under the name Diversity.
We obtain a comprehensive picture of the problem’s parameterized complexity and establish its
fixed-parameter tractability w.r.t. the solution size plus the power of the hypercube.

Given that several such FO-definable problems have been shown to be fixed-parameter tractable
on the considered graph class, one may ask whether fixed-parameter tractability could be extended
to capture all FO-definable problems. We answer this question in the negative by showing that FO
model checking on induced subgraphs of hypercubes is as difficult as FO model checking on general
graphs.

2012 ACM Subject Classification Theory of computation → Parameterized complexity and exact
algorithms

Keywords and phrases Independent Set, Powers of Hypercubes, Diversity, Parameterized Complexity,
Incomplete Data

Digital Object Identifier 10.4230/LIPIcs.IPEC.2023.16

Funding Robert Ganian: Robert Ganian acknowledges support from Project No. Y1329 of the
Austrian Science Fund (FWF).
Iyad Kanj: Iyad Kanj acknowledges support from DePaul University through URC grant 602061.
Sebastian Ordyniak: Project EP/V00252X/1 of the Engineering and Physical Sciences Research
Council (EPSRC).
Stefan Szeider: Stefan Szeider acknowledges support from Project No. P36420 of the Austrian
Science Fund (FWF).
1 Introduction

Recently, there has been an increasing interest in studying the parameterized complexity of clustering problems motivated by their applications in machine learning [2, 3, 4, 5, 6, 17, 19, 20, 25, 26, 29, 39, 40], particularly their applications to fundamental clustering problems [1, 28, 41, 44]. In many of these clustering problems, we are given a set of \( d \)-dimensional vectors over the Boolean/binary domain, where the vectors are regarded as rows of a matrix. It is worth noting that due to the applications of such problems in incomplete-data settings, a number of past works on the topic also studied settings where some of the entries in these vectors are unknown [30, 19, 20, 29, 17, 8, 9, 10, 21, 37]. The objective is to determine if these vectors (or, in the incomplete-data setting, their completions) satisfy some desirable clustering properties. Examples of such properties include admitting a partitioning into \( k \) clusters each of diameter (or radius) at most \( r \) (for some given \( k, r \in \mathbb{N} \)), or admitting a \( k \)-cluster of diameter (or radius) at most \( r \), where the distance under consideration is typically the Hamming distance [12, 16, 19, 20, 23, 32, 33, 34, 35]; here, a \( k \)-cluster of diameter \( r \) is a set of \( k \) points which have pairwise distance of at most \( r \).

As it turns out, many of these well-studied clustering problems can be formulated as classical graph problems on induced subgraphs of powers of the hypercube graph. For instance, finding a cluster of diameter at most \( r \in \mathbb{N} \), for a given \( r \), is equivalent to the \textsc{Clique} problem defined on the subgraph of the \( r \)-th power of the hypercube that is induced by the subset of hypercube vertices corresponding to the given input vectors. Similarly, partitioning the set of vectors into \( k \) clusters each of diameter at most \( r \), for some given \( r, k \in \mathbb{N} \), is equivalent to the partitioning into \( k \) cliques problem on the same graph class, whereas partitioning the set of vectors into clusters, each of radius at most \( r \) with respect to some vector in the set, is equivalent to the \( k \)-dominating set problem on the same graph class described above. We remark that, to the best of our knowledge, this graph class is not a subclass of commonly studied graph classes and has not been considered in previous works pertaining to algorithmic upper or lower bounds for graph-theoretic problems.

**Contribution.** In this paper, we study the parameterized complexity of another classical graph problem defined on induced subgraphs of powers of the hypercube: the \textsc{Independent Set} problem. In the context of data analytics, the problem arises when studying the “diversity” of a given set of vectors, a notion that can be viewed as the opposite of minimising the number of clusters in a cluster partitioning of the set of vectors (in fact, in the area of data analytics this is studied directly under the nomenclature diversity or dispersion [11, 31, 45]). More precisely, motivated by the aforementioned extensive interest in the analysis of incomplete data, we focus on the more general incomplete data setting. We refer to this problem as \textsc{Pow-Hyp-IS-Completion}: given a set of Boolean vectors with some missing entries and integers \( k \) and \( r \), the goal is to complete the missing entries so that the resulting set of vectors contains a subset \( S \) of \( k \) vectors such that the Hamming distance between each pair is at least \( r + 1 \) (or to correctly determine that such a set does not exist).

The main contribution of this paper is a complete characterisation of the parameterized complexity of \textsc{Pow-Hyp-IS-Completion} w.r.t. the two parameters \( k \) and \( r \): we provide a fixed-parameter algorithm for \textsc{Pow-Hyp-IS-Completion} when parameterized by \( k + r \), and complement this positive result with intractability results for the cases where any of these two parameters is dropped. In particular, we show that the problem is \textsc{NP}-complete already for \( r = 2 \) — that is, the problem is \textsc{paraNP}-hard parameterized by \( r \), and \textsc{W[1]}-hard parameterized by \( k \) alone. Interestingly, the \textsc{FPT} result shows that the parameterized complexity of the
problem is independent of any restrictions on the number or the structure of the missing entries in the input vectors – contrasting many of the previous results on clustering incomplete data \[30, 19, 20, 29\]. We remark that even the fixed-parameter tractability of the problem in the complete data setting (i.e., where all entries are known) is non-obvious, but follows as an immediate corollary of our result.

For our final contribution, we revisit the observation that several of the complete-data clustering problems recently considered in the literature (e.g., see [19, 20]) reduce to well-known graph problems on the class of induced subgraphs of powers of the hypercube. Since it was shown that all of these graph problems are fixed-parameter tractable when restricted to this graph class and the graph problems are expressible in First Order Logic (FO), a natural question to ask is whether these FPT results can be generalised to any graph problem expressible in FO logic. We resolve this question in the negative.

**Related Work.** The problem of computing the diversity of a data set, which forms the underpinning of our study of \textsc{Pow-Hyp-IS-Completion}, has been studied in a variety of different contexts and settings. For instance, Ceccarello, Pietracaprina, Pucci and Upfal studied approximation algorithms for the problem [11]. Gawrychowski, Krasnopolsky, Mozes, and Weimann obtained a linear-time algorithm for the problem when the data set is represented as a tree [31], improving upon the previous polynomial-time algorithm of Bhattacharya and Houle [7]. Sacharidis, Mehta, Skoutas, Patroumpas and Viosard provided heuristics for dynamic versions of the problem [45].

More broadly, there is extensive work on problems arising in the context of incomplete data. Hermelin and Rozenberg [38] studied the \textsc{Closest String with Wildcards} problem, which can be seen as the problem of finding a data completion and a center to a minimum-radius cluster containing all the data points. Koana, Froese and Niedermeier [39] recently revisited the earlier work of Hermelin and Rozenberg [38] and obtained, among other results, a fixed-parameter algorithm for that problem parameterized by the radius plus the maximum number of missing entries per row; see also the related work of the same authors [40]. Eiben et al. considered a number of different clustering problems in the presence of incomplete data [18, 19], and a subset of these authors previously investigated the fundamental \textsc{Matrix Completion} problem in the same setting [30]. The parameterized complexity of \textsc{k-means} clustering on incomplete data was investigated by Eiben et al. [17] and Ganian et al. [29].

**2 Preliminaries**

**Problem Terminology and Definition**

Let \( \vec{a} \) and \( \vec{b} \) be two vectors in \( \{0, 1, \square\}^d \), where \( \square \) is used to represent coordinates whose value is unknown (i.e., missing entries). We denote by \( \Delta(\vec{a}, \vec{b}) \) the set of coordinates in which \( \vec{a} \) and \( \vec{b} \) are guaranteed to differ, i.e., \( \Delta(\vec{a}, \vec{b}) = \{ i \mid (\vec{a}[i] = 1 \land \vec{b}[i] = 0) \lor (\vec{a}[i] = 0 \land \vec{b}[i] = 1) \} \), and we denote by \( \delta(\vec{a}, \vec{b}) \) the *Hamming distance* between \( \vec{a} \) and \( \vec{b} \) measured only between known entries, i.e., \( |\Delta(\vec{a}, \vec{b})| \). Moreover, for a subset \( D' \subseteq [d] \) of coordinates, we denote by \( \vec{a}[D'] \) the vector \( \vec{a} \) restricted to the coordinates in \( D' \).

Let \( M \subseteq \{0, 1, \square\}^d \) and let \( [d] = \{1, \ldots, d\} \). For a vector \( \vec{a} \in M \) and \( t \in \mathbb{N} \), we denote by \( N_t(\vec{a}) \) the \( t \)-Hamming *neighbourhood* of \( \vec{a} \), i.e., the set \( \{ \vec{b} \in M \mid \delta(\vec{a}, \vec{b}) \leq t \} \) and by \( N_t(M) \) the set \( \bigcup_{\vec{a} \in M} N_t(\vec{a}) \). We say that \( M^t \subseteq \{0, 1\}^d \) is a *completion* of \( M \subseteq \{0, 1, \square\}^d \) if there is a bijection \( \alpha : M \to M^t \) such that for all \( \vec{a} \in M \) and all \( i \in [d] \) it holds that either \( \vec{a}[i] = \square \) or \( \alpha(\vec{a})[i] = \vec{a}[i] \).

We now proceed to give the formal definition of the problem under consideration:
A parameterized problem is \( \text{fixed-parameter tractable} \) (FPT) \([24, 14, 13]\), if there is an algorithm, called an \( \text{FPT-algorithm} \), that decides whether an input \((I, \kappa)\) is a member of \(Q\) in time \(f(\kappa) \cdot |I|^{O(1)}\), where \(f\) is a computable function and \(|I|\) is the input instance size. The class FPT denotes the class of all fixed-parameter tractable parameterized problems.

A parameterized problem \(Q\) is \( \text{FPT-reducible} \) to a parameterized problem \(Q'\) if there is an algorithm, called an \( \text{FPT-reduction} \), that transforms each instance \((I, \kappa)\) of \(Q\) into an instance \((I', \kappa')\) of \(Q'\) in time \(f(\kappa) \cdot |I|^{O(1)}\), such that \(\kappa' \leq g(\kappa)\) and \((I, \kappa) \in Q\) if and only if \((I', \kappa') \in Q'\), where \(f\) and \(g\) are computable functions. Based on the notion of FPT-reducibility, a hierarchy of parameterized complexity, the \( \text{W-hierarchy} = \{W[i] : i \geq 0\} \), where \(W[t] \subseteq W[t+1]\) for all \(t \geq 0\), has been introduced, in which the 0-th level \(W[0]\) is the class \(\text{FPT}\). The notions of hardness and completeness have been defined for each level \(W[i]\) of the \( \text{W-hierarchy} \) for \(i \geq 1 \) \([14, 13]\). It is commonly believed that \(W[1] \neq \text{FPT}\) \([see [14, 13]]\). The \(W[1]\)-hardness has served as the main working hypothesis of fixed-parameter intractability.

A problem is \( \text{paraNP-hard} \) if it is \(\text{NP-hard}\) for a constant value of the parameter \([24]\).

**Sunflowers**

A \( \text{sunflower} \) in a set family \( \mathcal{F} \) is a subset \( \mathcal{F}' \subseteq \mathcal{F} \) such that all pairs of elements in \( \mathcal{F}' \) have the same intersection.

\( \blacktriangleright \text{Lemma 1} \) \([22, 24]\). \text{Let} \( \mathcal{F} \) \text{be a family of subsets of a universe} \( U \), \text{each of cardinality exactly} \( b \), \text{and let} \( a \in \mathbb{N} \). \text{If} \(|\mathcal{F}| \geq b(a-1)^b\), \text{then} \( \mathcal{F} \) \text{contains a sunflower} \( \mathcal{F}' \) \text{of cardinality at least} \( a \). \text{Moreover,} \( \mathcal{F}' \) \text{can be computed in time polynomial in} \(|\mathcal{F}|\).

3 \hspace{1em} \textbf{The Parameterized Complexity of} \hspace{1em} \textbf{Pow-Hyp-IS-Completion}

Our aim for \textbf{Pow-Hyp-IS-Completion} is to establish fixed-parameter tractability parameterized by \( k + r \) \((i.e., \text{regardless of the structure or number of missing entries})\). As our first step, we show that all rows in an arbitrary instance \((M, k, r)\) can be, \(w.l.o.g.,\) assumed to contain at most \( O(k \cdot r) \) many \( \Box \)'s.
Next, we observe that if $M$ is sufficiently large and the $r$-Hamming neighbourhood of each vector is upper-bounded by a function of $k + r$, then – since the number of $\square$’s is bounded – $(M, k, r)$ is a YES-instance. The argument here is analogous to the classical argument showing that INDEPENDENT SET is trivial on large bounded-degree graphs.

On a high level, we would now like to find and remove an “irrelevant vector” from $M$ – since here the number of $\square$’s on every row is bounded, any instance reduced in this way to only contain a bounded number of vectors can be solved via a brute-force fixed-parameter procedure. However, finding an irrelevant vector is rather challenging, primarily because the occurrence of $\square$’s is not restricted. Instead, we develop a more powerful set representation $\mathcal{F}'$ for vectors in the instance which also uses elements to keep track of the presence of $\square$’s in the neighbours of $\vec{v}$. We can then apply the Sunflower Lemma to find a sufficiently-large sunflower in $\mathcal{F}'$, and in the core of the proof we argue that (1) such a sunflower consists of at most a bounded number of “important petals” (which can be identified in polynomial time), and (2) any petal that is not important represents an irrelevant vector.

### 3.1 Dealing with Unstructured Missing Data

In this subsection, we design an algorithm for POW-HYP-IS-COMPLETION which remains efficient even when the number and placement of unknown entries is not explicitly restricted on the input.

We begin with a simple lemma that allows us to deal with vectors (i.e., rows) with a large number of missing entries. For brevity, let a $k$-diversity set be a set containing $k$ vectors which have pairwise Hamming distance at least $r + 1$.

**Lemma 2.** Let $\mathcal{I} = (M, k, r)$ be an instance of POW-HYP-IS-COMPLETION where $k \geq 1$ and let $\vec{v} \in M$ be a vector containing more than $(k - 1) \cdot (r + 1)$-many $\square$’s. Then $\mathcal{I}$ is a YES-instance if and only if $\mathcal{I}' = (M \setminus \{\vec{v}\}, k - 1, r)$ is a YES-instance. Moreover, a completion and $k$-diversity set for $\mathcal{I}$ can be computed from a completion and $(k - 1)$-diversity set for $\mathcal{I}'$ in linear time.

**Proof.** The forward direction is trivial: for any completion $M^*$ of $M$ and $k$-diversity set $S$ in $M^*$, we can obtain a $(k - 1)$-diversity set and completion for $\mathcal{I}'$ by simply removing $\vec{v}$ from $M^*$ and $S$.

For the backward direction, consider a completion $M''$ of $M' = M \setminus \vec{v}$ and a $(k - 1)$-diversity set $S = \{s_1, \ldots, s_{k-1}\}$ in $M''$. Let us choose an arbitrary set $C$ of $(k - 1) \cdot (r + 1)$ coordinates in $\vec{v}$ that all contain $\square$, and let us then partition $C$ into $k$-many subsets $\alpha_1, \ldots, \alpha_k$ each containing precisely $r + 1$ coordinates. Now consider the vector $\vec{v}^*$ obtained from $\vec{v}$ as follows:

1. For each $i \in [k - 1]$ and every coordinate $j \in \alpha_i$, set $\vec{v}^*[j]$ to the opposite value of $s_i[j]$ (i.e., $\vec{v}^*[j] = 1$ if and only if $s_i[j] = 0$);

2. For every other coordinate $j$ of $\vec{v}$, we set $\vec{v}^*[j] = \vec{v}[j]$ if $\vec{v}[j] \neq \square$ and $\vec{v}^*[j] = 0$ otherwise.

Clearly, $M^* = M'' \cup \{\vec{v}^*\}$ is a completion of $M$. Moreover, since $\vec{v}^*$ differs from each vector in $S$ in at least $r + 1$ coordinates, $S \cup \{\vec{v}^*\}$ is a $k$-diversity set in $M^*$.

Next, we show that instances which are sufficiently large and where each vector only “interferes with” a bounded number of other vectors are easy to solve. For ease of presentation, let $\zeta(k, r, t) = 3^{(k - 1)(r + 1)} \cdot t! \cdot \left( (k - 1) \cdot (3(k - 1) \cdot (r + 1) + r + t) \right)$ be the exact meaning of “sufficiently large” in this case.
16:6 From Data Completion to Problems on Hypercubes

Lemma 3. Let $\mathcal{I} = (M, k, r)$ be an instance of POW-HYP-IS-Completion. If $|M| \geq k \cdot r \cdot \zeta(k, r, r)$ and $|N_i(\bar{v})| \leq \zeta(k, r, r)$ for every $\bar{v} \in M$ and $t \leq r$, then a $k$-diversity set in $\mathcal{I}$ can be found in polynomial time.

Proof. One can find a solution to $\mathcal{I}$ by iterating the following greedy procedure $k$ times: choose an arbitrary vector $\bar{v}$, add it into a solution, and delete all other vectors with Hamming distance at most $r$ from $\bar{v}$. By the bound on $|N_i(\bar{v})|$, each choice of $\bar{v}$ will only lead to the deletion of at most $r \cdot \zeta(k, r, r)$ vectors from $M$. Moreover, since $\delta$ measures the Hamming distance only between known entries, any completion of the missing entries can only increase (and never decrease) the Hamming distance between vectors. Hence, the size of $M$ together with the bounded size of the Hamming neighbourhood of $\bar{v}$ guarantee that this procedure will find a solution of cardinality $k$ in $\mathcal{I}$ which will remain valid for every completion of $M$. ▶

We can now move on to the main part of the proof: a procedure which either outputs a solution outright or finds an irrelevant vector.

Lemma 4. Let $\mathcal{I} = (M, k, r)$ be an instance of POW-HYP-IS-Completion such that $|N_i(\bar{v})| \geq \zeta(k, r, r)$ for some vector $\bar{v} \in M$ and $t \leq r$ and such that each vector in $M$ contains at most $(k-1) \cdot (r+1)$ $\Box$’s. There is a polynomial-time procedure that finds a vector $\bar{f} \in M$ satisfying the following properties:

1. $(M, k, r)$ is a YES-instance if and only if $\mathcal{I}' = (M \setminus \{\bar{f}\}, k, r)$ is a YES-instance, and
2. A completion and diversity set for $\mathcal{I}$ can be computed from a solution and diversity set for $\mathcal{I}'$ in linear time.

Proof. We will begin by constructing a set system over the neighbourhood of $\bar{v}$. Let $Z = \{ z \in [d] \mid \bar{v}[z] = \Box \}$ be the set of coordinates where $\bar{v}$ is incomplete. Clearly, since $|N_i(\bar{v})| \geq 3(k-1)(r+1)t! \left( (k-1) \cdot (3(k-1)(r+1)+r+t) \right)^t$ and $|Z| \leq (k-1)(r+1)t!$, we can find a subset $N \subseteq N_i(\bar{v})$ of vectors whose cardinality is at least $t! \left( (k-1) \cdot (3(k-1)(r+1)+r+t) \right)^t$ such that all vectors in $N$ are the same on the coordinates in $Z$, i.e., $\forall \bar{x}, \bar{y} \in N : \forall z \in Z : \bar{x}[z] = \bar{y}[z]$.

Now, let $F$ be a set containing 2 elements for each coordinate $j \in [d] \setminus Z$ of vectors in $M$: the element $\Box_j$, and the element $D_j$. We construct a set system $\mathcal{F}$ over $F$ as follows: for each vector $\bar{x} \in N$, we add a set $\bar{x}$ to $\mathcal{F}$ that contains:

1. $\Box_j$ if and only if $\bar{x}[j] = \Box$, and
2. $D_j$ if and only if $\bar{x}[j] \neq \Box_j$.

Observe that, since $\bar{x}$ contains at most $(k-1) \cdot (r+1)$ $\Box$’s by assumption and since $\bar{x}$ differs from $\bar{v}$ in at most $t$-many completed coordinates, every set in $\mathcal{F}$ has cardinality at most $(k-1) \cdot (r+1) + t$. This means we can apply Lemma 1 to find a sunflower $\mathcal{F}'$ in $\mathcal{F}$ of cardinality at least $(k-1) \cdot (3(k-1)(r+1)+r+t) + 1$; for ease of presentation, we will identify the elements of $\mathcal{F}'$ with the vectors they represent. Let $\bar{f}$ be an arbitrarily chosen vector from $\mathcal{F}'$: we claim that $\bar{f}$ satisfies the properties claimed in the lemma, and to complete the proof it suffices to establish this claim.

The backward direction is trivial: if $\mathcal{I}'$ is a YES-instance then clearly $\mathcal{I}$ is a YES-instance as well. It is also easy to observe that a completion and diversity set for $\mathcal{I}$ can be computed from a solution and diversity set for $\mathcal{I}'$ in linear time (adding a vector does not change the validity of a solution). What we need to show is that if $\mathcal{I}$ is a YES-instance, then so is $\mathcal{I}'$ (i.e., $(M \setminus \{\bar{f}\}, k, r)$); moreover, this final claim clearly holds if $\mathcal{I}$ admits a solution that does not contain $\bar{f}$.
So, assume that \( M \) admits a completion \( M^* \) which contains a \( k \)-diversity set \( S = \{ f, s_1, \ldots, s_k-1 \} \). Let \( C \) be the core of the sunflower \( F' \), and note that all vectors in \( F' \) have precisely the same content in the coordinates in \( C \).

Finding a replacement for \( f \). We would now like to argue that, for some completion which we will define later, \( F' \) contains a vector that can be used to replace \( f \) in the solution.

Let \( s'_i \in S \) be an arbitrary vector. First, let us consider the case that, in \( M \), \( s'_i \) differs from \( v \) in more than \( 3(k-1) \cdot (r+1) + r + t \) coordinates (i.e., \( v[j] \neq s'_i[j] \) in \( M \) for at least \( 3(k-1) \cdot (r+1) + r + t \) choices of \( j \)). Then every vector in \( F' \) will have Hamming distance greater than \( r \) from \( s'_i \) regardless of the completion.

Indeed, for every vector \( f' \in F' \) there are at most \( 3(k-1) \cdot (r+1) \) coordinates \( j \) such that at least one of \( v[j], s'_i[j], f' \), meaning that there are at least \( r + t \) other coordinates where \( v \) differs from \( s'_i \) and which are guaranteed to be complete – and since \( \delta(f', v) = t \), \( f' \) it must hold that \( \delta(f', v) > r \) (by the triangle inequality). Hence indeed every vector in \( F' \) must have distance at least \( r + 1 \) from \( s'_i \), and in this case we will create a set \( S_i = \emptyset \) (the meaning of this will become clear later).

Now, consider the converse case, i.e., that \( s'_i \) differs from \( v \) in at most \( 3(k-1) \cdot (r+1) + r + t \) coordinates. We may now extend the sunflower \( F' \) by adding a set representation of \( s'_i \), i.e., a set \( Q_i \) which contains \( \emptyset \), if and only if \( s'_i[j] = \emptyset \) and \( D_j \), if and only if \( s'_i[j] \neq v[i] \) (for all \( j \in [d] \setminus Z \)). Observe that \( |Q_i| \leq 3(k-1) \cdot (r+1) + r + t \), and in particular \( Q_i \cap C \) intersects with at most \( 3(k-1) \cdot (r+1) + r + t \) elements of \( F' \). Let \( S_i \) be the set of all such elements, i.e., elements of \( F' \) which have a non-empty intersection with \( Q_i \) outside of the core (formally, with \( Q_i \setminus C \)).

To conclude the proof, we will show that there is a completion \( M^{*\prime} \) of \( M' \) such that any arbitrarily chosen vector \( f' \) in the non-empty set \( F' \setminus \left( \{ f \} \cup \bigcup_{i \in [k-1]} S_i \right) \) can replace \( f \) in the \( k \)-diversity set \( S \).

Arguing Replaceability. Consider a new completion \( M^{*\prime} \) of \( M \setminus f \) obtained as follows:

- For each vector \( w \in F' \setminus S \), we complete
  1. the \( \emptyset \)'s in \( C \cup Z \) precisely in the same way as \( f \), and
  2. for every other \( \emptyset \) at coordinate \( j \), we set \( w[j] = -(v[j] - 1) \) (i.e., to the opposite of \( v \); recall that \( v[j] \neq \emptyset \) since \( j \not\in Z \));
- all other \( \emptyset \)'s in all other vectors in \( M \setminus f \) are completed in precisely the same way as in \( M^{*\prime} \).

Since \( M^{*\prime} \) precisely matches \( M^* \) on all vectors in \( S \setminus f \), it follows that \( S \setminus f \) is a \( (k-1) \)-diversity set in \( M^{*\prime} \). Moreover, consider for a contradiction that \( \delta(f', s'_i) \leq r \) for some \( s'_i \in S \) after completion, i.e., in \( M^{*\prime} \). Then clearly \( s'_i \) could not differ from \( v \) in more than \( 3(k-1) \cdot (r+1) + r + t \) coordinates in \( M' \), since – as we already argued – in this case every vector in \( F' \) will have Hamming distance greater than \( r \) from \( s'_i \) regardless of the completion.

Hence, we must be in the case where \( s'_i \) differed from \( v \) in at most \( 3(k-1) \cdot (r+1) + r + t \) coordinates in \( M' \). Now consider how \( \delta(f', s'_i) \) differs from \( \delta(f, s'_i) \). First of all, there is no difference between these two distances on the coordinates in \( Z \cup C \) due to our construction of \( M^{*\prime} \) and choice of \( N \). For the remaining coordinates, we will consider separately the set \( X \) of coordinates in the petals of \( f' \) (i.e., the set \( \{ j \in [d] \setminus (Z \cup C) \mid f'[j] \neq f[j] \lor f'[j] \neq v[j] \} \)), and the set \( Y = [d] \setminus (C \cup Z \cup X) \) of all remaining coordinates. It follows that \( \forall \emptyset | [j] = [j] = f'[j] \) for all coordinates \( j \in Y \), and hence there is no difference between the two distances on these coordinates either.
So, all that is left is to consider the difference between \( \delta(\vec f', \vec s_i) \) and \( \delta(\vec f, \vec s_i) \) on the coordinates in \( X \). Among these coordinates, \( \vec f \) can only differ from \( \vec s_i \) in at most \( t - |C| \) many coordinates – notably in the coordinates of its own petal – because the coordinates in the petal of \( \vec f' \) do not intersect with \( Q_s \). On the other hand, our construction guarantees that \( \vec f' \) differs from \( \vec s_i \) in at least \( t - |C| \) coordinates in \( X \); more precisely, on all coordinates in the petal of \( \vec f' \), since on these coordinates (1) \( \vec s_i \) is equal to \( \vec v \) and (2) \( \vec f' \) differs from \( \vec v \).

In summary, we conclude that \( \delta(\vec f', \vec s_i) \geq \delta(\vec f, \vec s_i) \) and hence \( (S \setminus \{\vec f\}) \cup \{\vec f'\} \) is a \( k \)-diversity set in \( M^n \), as claimed.

We can now establish our main result for Pow-Hyp-IS-Completion.

**Theorem 5.** Pow-Hyp-IS-Completion is fixed-parameter tractable parameterized by \( k + r \).

**Proof.** The algorithm proceeds as follows. Given an instance \( I = (M, k, r) \) of Pow-Hyp-IS-Completion, it first checks whether \( M \) contains a vector with more than \( (k - 1) \cdot (r + 1) \) \( \Box \)'s; if yes, it applies Lemma 2 and restarts on the reduced instance. Second, it checks whether \(|M| \geq k \cdot r \cdot \zeta(k, r, t)\); if not, it uses the fact that the number of \( \Box \)'s and the number of rows is bounded by a function of the parameter to find a completion and a \( k \)-diversity set in \( I \) (or determine that one does not exist) by brute force.

Third, it checks whether each vector \( \vec v \) satisfies \(|N_t(\vec v)| < \zeta(k, r, t)\) for every \( t \in [r] \); if yes, then it solves \( I \) by invoking Lemma 3. Otherwise, it invokes Lemma 4 to reduce the cardinality of \( M \) by \( 1 \) and restarts. If the algorithm eventually terminates with a “NO”, then we know that the initial input was a NO-instance; otherwise, it will output a solution which can be transformed into a solution for the original input by the used lemmas.

### 3.2 Lower Bounds

**Theorem 6.** Pow-Hyp-IS is NP-complete and \( W[1] \)-hard parameterized by \( k \).

**Proof.** We prove both NP-hardness and \( W[1] \)-hardness results by giving a polynomial-time FPT reduction from Independent Set (IS), which is \( W[1] \)-hard [14].

Let \( (G, k) \) be an instance of IS, where \( V(G) = \{v_1, \ldots, v_n\} \), and let \( m = E(G) \). Fix an arbitrary ordering \( \mathcal{O} = (e_1, \ldots, e_m) \) of the edges in \( E(G) \).

For each vertex \( v_i \in V(G) \), define a vector \( \vec a_i \in \{0, 1\}^m \) by setting \( \vec a_i[j] = 1 \) if \( v_i \) is incident to \( e_j \) and \( \vec a_i[j] = 0 \) otherwise. Now expand the set of coordinates of these vectors by adding to each of them \( n(n - 1) \) new coordinates, \( n - 1 \) coordinates for each \( v_i, i \in [n] \); we refer to the \( n - 1 \) (extra) coordinates of \( v_i \) as the “private” coordinates of \( v_i \). For each \( v_i, i \in [n] \), set \( n - 1 - \text{deg}(v_i) \) many coordinates among the private coordinates of \( v_i \) to 1, and all other new coordinates of \( v_i \) to 0. Let \( M = \{\vec a_i \mid i \in [n]\} \) be the set of expanded vectors, where \( \vec a_i \in \{0, 1\}^{n + n(n - 1)} \), for \( i \in [n] \). The reduction from IS to Pow-Hyp-IS produces the instance \( I = (M, k, 2n - 4) \) of Pow-Hyp-IS; clearly, this reduction is a polynomial-time FPT-reduction.

Observe that, for any two distinct vertices \( v_i, v_j \in V(G) \), \( \delta(\vec a_i, \vec a_j) = 2n - 2 \) if \( v_i \) and \( v_j \) are nonadjacent and \( \delta(\vec a_i, \vec a_j) = 2n - 4 \) if \( v_i \) and \( v_j \) are adjacent.

The proof that \( (G, k) \) is a Yes-instance of IS iff \( (M, k, 2n - 4) \) is a Yes-instance of Pow-Hyp-IS is now straightforward.

**Theorem 7.** Pow-Hyp-IS is NP-complete even when \( r = 2 \).
Proof. We reduce from the INDEPENDENT SET problem (which is NP-complete). Let \((G, k)\)
be an instance of INDEPENDENT SET and let \(G'\) be the graph obtained from \(G\) after subdividing every edge exactly twice. We first observe that \(G\) has an independent set of size at least \(k\) if and only if \(G'\) has an independent set of size at least \(|E(G)| + k\). This is because if \(I \subseteq V(G)\) is an independent set of \(G\), then we can add one of the subdivision vertices for every edge of \(G\) because \(I\) does not contain both endpoints of an edge. On the other hand, if \(I \subseteq V(G')\) is an independent set of \(G'\), then we can assume without loss of generality that \(I\) does not contain both endpoints of an edge in \(G\) because we could easily transform \(I\) into an independent set of the same size by replacing one of the endpoints of such an edge with a subdivided vertex.

Next we construct an instance \(I = (M, |E(G)| + k, 2)\) of POW-HYP-IS in polynomial-time such that \(G'\) has an independent set of size at least \(|E(G)| - k\) if and only if \(I\) is a Yes-instance. We set \(d = 2|V(G)|\) and obtain \(M\) as follows. Let \(V(G) = \{v_1, \ldots, v_n\}\). For every \(v_i \in V(G)\), we add the vector \(\vec{v}_i\) that is 1 at the two coordinates \(i\) and \(i + 1\) and otherwise 0. Moreover, for every \(e = v_iv_j \in E(G)\), we add the vector \(e^1\) that is 1 at the coordinates \(i\), \(i + 1\), and \(j\) and the vector \(e^2\) that is 1 at the coordinates \(j\), \(j + 1\), and \(i\). This completes the construction of \(I\). The equivalence now follows because two vectors in \(M\) have distance at most \(r = 2\) if and only if their corresponding vertices in \(G'\) are adjacent; here \(e^1\) and \(e^2\) correspond to the two subdivision vertices on the edge \(e\).

4 On Graph Problems on Induced Subgraphs of the Hypercubes

In this section, we discuss the implications of the results in the previous section for fundamental
problems defined on induced subgraphs of powers of the hypercube graph.

In particular, the \(d\)-dimensional hypercube graph is the graph \(Q_d\) whose vertex set is the
set of all Boolean \(d\)-dimensional vectors, and two vertices are adjacent if and only if their
two vectors differ in precisely 1 coordinate. We can then define the class \(Q_d\) as the class of
all graphs that are induced subgraphs of the \(r\)-th power of \(Q_d\). We note that, in line with
the commonly used definition of hypercube graphs [15, 27], we consider the vertices in \(Q_d\) to
be vectors and hence every graph \(G \in Q_d\) contains an explicit characterisation of its vertices
as vectors.

In this setting, it is straightforward to observe that POW-HYP-IS is precisely the IN-
DEPENDENT SET problem on \(Q_d\). Moreover, the clustering problems IN-CLUSTERING,
DIAM-CLUSTERING, and LARGE DIAM-CLUSTER considered in [19, 20] are precisely the
DOMINATING SET, PARTITION INTO CLIQUES, and CLIQUE problems, respectively, on
\(Q_d\). Therefore, all the upper and lower bound results derived in this paper and in [19, 20]
pertaining to these clustering problems hold true for their corresponding graph problems
on \(Q_d\).

\(\triangleright\) Corollary 8. Given \(r, d, k \in \mathbb{N}\) and a graph \(G \in Q_d\), determining whether \(G\) has a:
- dominating set of size \(k\) is FPT parameterized by \(k + r\);
- partition into \(k\) cliques is FPT parameterized by \(k + r\);
- independent set of size \(k\) is FPT parameterized by \(k + r\);
- clique of size \(k\) is FPT parameterized by \(r\).

We note that all the tractability results outlined in Corollary 8 are tight, which follows from
the lower-bound results obtained in Section 3.2 and in [19, 20], in the sense that
dropping any parameter from our parameterizations leads to an intractable problem.
Observing that three of the graph properties in the problems discussed above are expressible in First Order Logic (FO) and result in FO formulas whose length is a function of the parameter \( k \), an interesting question that ensues from the above discussion is whether these positive results can be extended to the generic problem of First-Order Model Checking [43, 36], formalised below. We will show next that the answer to this question is negative – and, in fact, remains negative even when we restrict ourselves to induced subgraphs of hypercubes (i.e., for \( r = 1 \)).

We denote by \( \text{FO-Model-Checking} \) the general FO Model Checking problem on graphs, i.e., \( \text{C-FO-Model-Checking} \) with \( \text{C} \) being the class of all graphs.

Lemma 9. Let \( H \) be an arbitrary graph. There is a graph \( G \in \mathcal{Q}_d^1 \) such that \( G \) is isomorphic to the graph \( H' \) obtained from \( H \) after subdividing every edge of \( H \) exactly once and attaching a leaf to every vertex resulting from a subdivision. Moreover, \( G \) can be computed from \( H \) in polynomial time.

**Proof.** Let \( n = |V(H)| \) and \( m = |E(H)| \). To prove the lemma, we construct a matrix representation \( M \in \{0,1\}^{n+m} \) of \( H' \) which has one row (vector) for every vertex in \( H \) and where two vertices in \( H' \) are adjacent if and only if their corresponding rows in \( M \) have Hamming distance at most 1. Let \( v_1, \ldots, v_n \) be an arbitrary ordering of the vertices of \( H \), and \( e_1, \ldots, e_m \) be an arbitrary ordering of its edges. Then, \( M \) contains one row \( r_i \) for every \( i \in [n] \) that is 1 at its \( i \)-th entry and 0 at all other entries. Moreover, for every edge \( e_i = \{v_i, v_j\} \in E(H) \), \( M \) contains the following two rows:

- the row \( r_{e_i} \) (corresponding to the degree-3 vertex in \( H' \) obtained from \( e \)) that is 1 at the \( i \)-th and \( j \)-th entries, and 0 at all other entries; and
- the row \( r_{e_i}^0 \) (corresponding to the leaf in \( H' \) obtained from \( e \)) that is 1 at the \( i \)-th, \( j \)-th, and \( (n+1) \)-th entries, and 0 at all other entries.

This completes the construction of \( M \). Clearly, two rows in \( M \) have Hamming distance at most one if and only if their corresponding vertices in \( H' \) are adjacent, as required. ⊳

Theorem 10. \( \text{Q-FO-model-checking} \) is \( \text{W}[t] \)-hard for every \( t \in \mathbb{N}^* \).

**Proof.** We give a parameterized reduction from \( \text{FO Model Checking} \), which is \( \text{W}[t] \)-hard for every \( t \in \mathbb{N}^* \). Let \( I := (\Phi, H) \) be an instance of \( \text{FO Model Checking} \). We will show the theorem by constructing the equivalent instance \( I' := (\Phi', G) \) such that \( G \in \mathcal{Q}_d^1 \) and \( |\Phi| \leq f(|\Phi'|) \) for some computable function \( f \) and value \( d \) that is polynomially bounded in the input size. \( G \) is obtained from \( H \) in the same manner as in Lemma 9. Moreover, \( \Phi' \) is obtained from \( \Phi \) as follows:

- Let \( \phi_V(x) \) be the formula that holds for a variable \( x \) if and only if \( x \) corresponds to one of the original vertices in \( G \), i.e., \( \phi_V(x) := \forall y E(x, y) \exists z \neq x \land E(y, z) \); replace every subformula of the form \( \exists x \phi \) (for some variable \( x \) and some subformula \( \phi \) of \( \Phi \)) with the formula \( \exists x \phi_V(x) \land \phi \); and replace every subformula of the form \( \forall x \phi \) (for some variable \( x \) and some subformula \( \phi \) of \( \Phi \)) with the formula \( \forall x \phi_V(x) \rightarrow \phi \); and replace every atom \( E(x, y) \), where \( E \) is the adjacency predicate and \( x \) and \( y \) are variables, with the formula \( \exists s E(x, s) \land E(s, y) \land x \neq y \).

It is straightforward now to show that \( H \models \Phi \) if and only if \( G \models \Phi' \), and that \( |\Phi'| \leq 20|\Phi| \). Moreover, because of Lemma 9, \( G' \in \mathcal{Q}_d^1 \), as required. ⊳
5 Conclusion

In this paper, we studied the parameterized complexity of the classical Independent Set problem on induced subgraphs of powers of hypercubes, but with the additional complication that the “positions” of the vertices in the hypercube representation may be partially unknown. We considered the two most natural parameters for the problem: the size $k$ of the independent set and the power $r$ of the hypercube, and provided a complete characterisation of the problem’s complexity w.r.t. $k$ and $r$. We also performed a meta-investigation of the parameterized complexity of graph problems on this graph class that are expressible in FO logic and showed the existence of such problems that are parameterized intractable.

A natural future direction of our work is to study the parameterized complexity of other graph problems on this class, in particular those that have applications in clustering. One famous open problem that comes to mind is the $p$-center problem [16, 32]. The problem can be formulated similarly to the above setting, with the exception of allowing the selection of vertices to be from the whole hypercube, as opposed to restricting them to the input subgraph. In particular, the well-known $p$-centers problem reduces to the $k$-dominating set problem in the $r$-th power of the hypercube graph, but where the $k$ vertices in the dominating set are not restricted to the input subgraph, but can be chosen from $Q_d$. This problem was shown to be FPT parameterized by $k + r$ [20]. An intriguing NP-hard restriction of the problem is the problem slice corresponding to $p = 1$, or what is known as the 1-center problem, or equivalently, the Closest String problem [32, 42]. The parameterized complexity of the problem parameterized by each of $k$ and $r$ alone remain important open questions.

References

From Data Completion to Problems on Hypercubes


From Data Completion to Problems on Hypercubes

