Faster Property Testers in a Variation of the Bounded Degree Model

ISOLDE ADLER and POLLY FAHEY, University of Bamberg, Germany

Property testing algorithms are highly efficient algorithms that come with probabilistic accuracy guarantees. For a property P, the goal is to distinguish inputs that have P from those that are *far* from having P with high probability correctly, by querying only a small number of local parts of the input. In property testing on graphs, the *distance* is measured by the number of edge modifications (additions or deletions) that are necessary to transform a graph into one with property P. Much research has focused on the *query complexity* of such algorithms, i. e., the number of queries the algorithm makes to the input, but in view of applications, the *running time* of the algorithm is equally relevant.

In (Adler, Harwath, STACS 2018), a natural extension of the bounded degree graph model of property testing to relational databases of bounded degree was introduced, and it was shown that on databases of bounded degree and bounded tree-width, every property that is expressible in monadic second-order logic with counting (CMSO) is testable with constant query complexity and *sublinear* running time. It remains open whether this can be improved to constant running time.

In this article we introduce a new model, which is based on the bounded degree model, but the distance measure allows both edge (tuple) modifications and vertex (element) modifications. We show that every property that is testable in the classical model is testable in our model with the same query complexity and running time, but the converse is not true. Our main theorem shows that on databases of bounded degree and bounded tree-width, every property that is expressible in CMSO is testable with constant query complexity and *constant* running time in the new model. Our proof methods include the semilinearity of the neighborhood histograms of databases having the property and a result by Alon (Proposition 19.10 in Lovász, Large networks and graph limits, 2012) that states that for every bounded degree graph $\mathcal{G}$ there exists a constant size graph $\mathcal{H}$ that has a similar neighborhood distribution to $\mathcal{G}$.

It can be derived from a result in (Benjamini et al., Advances in Mathematics 2010) that hyperfinite hereditary properties are testable with constant query complexity and constant running time in the classical model (and hence in the new model). Using our methods, we give an alternative proof that hyperfinite hereditary properties are testable with constant query complexity and constant running time in the new model.

We argue that our model is natural and our meta-theorem showing constant-time CMSO testability supports this.

CCS Concepts: • **Theory of computation** → **Streaming, sublinear and near linear time algorithms**; **Database query processing and optimization (theory);**

Additional Key Words and Phrases: Constant time algorithms, logic and databases, property testing, bounded degree model

An extended abstract of this article appeared in the proceedings of FSTTCS 2020.

Authors' addresses: I. Adler, Otto-Friedrich-Universität Bamberg, Fakultät Wirtschaftsinformatik und Angewandte Informatik, Lehrstuhl für Algorithmen und Komplexitätstheorie, 96045 Bamberg, Germany; email: isolde.adler@uni-bamberg.de; P. Fahey, University of Leeds, School of Computing, Leeds, UK, LS2 9JT; email: mm11pf@leeds.ac.uk.

1529-3785/2023/05-ART25 $15.00

https://doi.org/10.1145/3584948

ACM Reference format:
Isolde Adler and Polly Fahey. 2023. Faster Property Testers in a Variation of the Bounded Degree Model. *ACM Trans. Comput. Logic* 24, 3, Article 25 (May 2023), 24 pages.
https://doi.org/10.1145/3584948

1 INTRODUCTION

Extracting information from large amounts of data and understanding its global structure can be an immensely challenging and time-consuming task. When the input data is huge, many traditionally "efficient" algorithms are no longer practical. The framework of property testing aims at addressing this problem. Property testing algorithms (*testers*, for short) are given oracle access to the inputs, and their goal is to distinguish between inputs that have a given property **P** or are structurally *far* from having **P** with high probability correctly. This can be seen as a relaxation of the classical yes/no decision problem for **P**. Testers make these decisions by exploring only a small number of local parts of the input that are randomly chosen. They come with probabilistic guarantees on the quality of the answer. Typically, only a constant number of small local parts are explored and the algorithms often run in constant or sublinear time. This speedup in running time, while sacrificing some accuracy, can be crucial for dealing with large inputs. In particular, it can be useful for a quick exploration of newly obtained data (e.g., biological networks). Based on the outcome of the exploration, a decision can then be taken whether to use a more time-consuming exact algorithm in a second step.

A *property* is simply an isomorphism-closed class of graphs or relational databases. For example, each Boolean database query q defines a property $\mathbf{P}_q$, the class of all databases satisfying q. In the bounded degree graph model [18], a uniform upper bound d on the degree of the graphs is assumed. For a small $\epsilon \in (0, 1]$, two graphs $\mathcal{G}$ and $\mathcal{H}$, both on n vertices, are *ϵ-close*, if at most ϵdn edge modifications (deletions or insertions in $\mathcal{G}$ or $\mathcal{H}$) are necessary to make $\mathcal{G}$ and $\mathcal{H}$ isomorphic. If $\mathcal{G}$ and $\mathcal{H}$ are not *ϵ-close*, then they are called *ϵ-far*. A graph $\mathcal{G}$ is called *ϵ-close* to a property **P** if $\mathcal{G}$ is *ϵ-close* to a member of **P**, and $\mathcal{G}$ is *ϵ-far* from **P** otherwise. The natural generalization of this model to relational databases of bounded degree (where a database has degree at most d if each element in its domain appears in at most d tuples) was studied in [1], where two databases $\mathcal{D}$ and $\mathcal{D}'$, both with n elements in the domain, are *ϵ-close*, if at most ϵdn tuple modifications (deletions from relations or insertions to relations) are necessary to make $\mathcal{D}$ and $\mathcal{D}'$ isomorphic, and $\mathcal{D}$ and $\mathcal{D}'$ are *ϵ-far* otherwise. We call this model for bounded degree relational databases the BDRD model.

Our contributions. In this article we propose a new model for property testing on bounded degree relational databases, which we call the $\text{BDRD}_{+/-}$ model, with a distance measure that allows both tuple deletions and insertions, and *deletion and insertion of elements of the domain*. On graphs, this translates to edge insertions and deletions, and *vertex insertions and deletions*. We argue that this yields a natural distance measure. Indeed, take any (sufficiently large) graph $\mathcal{G}$, and let $\mathcal{H}$ be obtained from $\mathcal{G}$ by adding an isolated vertex. Then $\mathcal{G}$ and $\mathcal{H}$ are ϵ-far for every $\epsilon \in (0, 1]$ under the classical distance measure, although they only differ in one vertex. In contrast, our distance measure allows for a small number of vertex modifications. While comparing graphs on different numbers of vertices by adding isolated vertices was done implicitly as part of the study of the testability of outerplanar graphs [5], to the best of our knowledge, such a distance measure has not been considered before as part of a model in property testing, which seems surprising to us.

Formally, in the $\text{BDRD}_{+/-}$ model, two databases $\mathcal{D}$ and $\mathcal{D}'$ are *ϵ-close* if they can be made isomorphic by at most *ϵdn modifications*, where a modification is either (1) removing a tuple from a relation, (2) inserting a tuple into a relation, (3) removing an element from the domain (and, as a consequence, any tuple containing that element is removed), or (4) inserting an element into the domain. Here n is the minimum of the sizes of the domains of $\mathcal{D}$ and $\mathcal{D}'$. In Section 3 we give the full details of our model. We note that the BDRD$_{+/-}$ model differs from the BDRD model only in the choice of the distance measure. While we work in the setting of relational databases, we would like to emphasize that our results carry over to (undirected and directed) graphs, as these can be seen as special instances of relational databases.

It is known that in the bounded degree graph model, every minor-closed property is testable [7], and, more generally, every hyperfinite graph property is testable [25] with constant query complexity. However, no bound on the running time can be obtained in these general settings. Indeed, there exist hyperfinite properties (of edgeless graphs) that are uncomputable. In [1], Adler and Harwath ask which conditions guarantee both low query complexity *and* efficient running time. They prove a meta-theorem stating that, on classes of databases (or graphs) of bounded degree and bounded tree-width, every property that can be expressed by a sentence of **monadic second-order logic with counting (CMSO)** is testable with *constant* query complexity and *polylogarithmic* running time in the BDRD model. Treating many algorithmic problems simultaneously, this can be seen as an algorithmic *meta-theorem* within the line of research inspired by Courcelle's famous theorem [10] that states that each property of relational databases that is definable in CMSO is decidable in linear time on relational databases of bounded tree-width. CMSO extends **first-order logic (FO),** and hence properties expressible in FO (e.g., subgraph/sub-database freeness) are also expressible in CMSO. Other examples of graph properties expressible in CMSO include bipartiteness, colorability, even-hole-freeness, and Hamiltonicity. Rigidity (i.e., the absence of a non-trivial automorphism) cannot be expressed in CMSO (cf. [11] for more details).

Our main theorem (Theorem 21) shows that in the BDRD$_{+/-}$ model, on classes of databases (or graphs) of bounded degree and bounded tree-width, every property that can be expressed by a sentence of CMSO is testable with *constant* query complexity and *constant* running time. The question whether constant running time can also be achieved in the BDRD model remains open.

We show that the BDRD$_{+/-}$ model is in fact stronger than the BDRD model: any property testable in the BDRD model is also testable in the BDRD$_{+/-}$ model with the same query complexity and running time (Lemma 4), but there are examples that show that the converse is not true (Lemma 6).

We also discuss the constant time testability of hyperfinite hereditary properties in the BDRD and BDRD$_{+/-}$ models (Theorems 27 and 28, respectively). To the best of our knowledge it has not been shown explicitly that hyperfinite hereditary properties are uniformly testable in constant time (in the bounded degree graph or BDRD models). In [7] it is proved that every monotone hyperfinite property is constant time testable in the bounded degree graph model. In [12] the authors prove that hereditary properties are testable in constant time on classes of non-expanding hereditary properties (which include hyperfinite hereditary classes) in the bounded degree graph model. We sketch a proof that hyperfinite hereditary properties are uniformly testable in constant time in the BDRD model (and hence the BDRD$_{+/-}$ model) using methods similar to [7] and [12]. We then give an alternative proof showing that hyperfinite hereditary properties are uniformly testable in constant time in the BDRD$_{+/-}$ model using different techniques, similar to those used for our main theorem (Theorem 21).

In the future, it would be interesting to obtain a characterization of the properties that are (efficiently) testable in the BDRD$_{+/-}$ model.

Our techniques. We assume a fixed upper bound on the degree of all databases. For proving our main theorem, we give a general condition under which properties are testable in constant time in the BDRD$_{+/-}$ model. To describe this condition let us first briefly introduce some terminology.

A property **P** is *hyperfinite* on a class of databases **C** if every database in **P** can be partitioned into connected components of constant size by removing only a constant fraction of the tuples such that the resulting partitioned database is in **C**. For $r \in \mathbb{N}$ and an element a in the domain of a database $\mathcal{D}$, the *r-neighborhood type* of a in $\mathcal{D}$ is the isomorphism type of the sub-database of $\mathcal{D}$ induced by all elements that are at distance at most r from a in the underlying graph of $\mathcal{D}$, expanded by a. The *r-histogram* of a bounded degree database $\mathcal{D}$, denoted by $\mathrm{h}_r(\mathcal{D})$, is a vector indexed by the r-neighborhood types, where the component corresponding to the r-neighborhood type τ contains the number of elements in $\mathcal{D}$ that realize τ. The *r-neighborhood distribution* of $\mathcal{D}$ is the vector $\mathrm{h}_r(\mathcal{D})/n$, where $\mathcal{D}$ is on n elements. We show that for any property **P** and input class **C**, if **P** is hyperfinite on **C** and the set of r-histograms of the databases in **P** are semilinear, then **P** is testable on **C** in constant time (Theorem 20). As a corollary we then obtain our main theorem, that every property definable by a CMSO sentence is testable on the class of databases with bounded degree and bounded tree-width in constant time (Theorem 21).

In addition, we prove a more general version of Theorem 20. We show that for any property **P** and input class **C**, if **P** is hyperfinite on **C** and the set of r-histograms of the databases in **P** are close to being semilinear, then **P** is testable on **C** in constant time (Theorem 25). We show that hyperfinite hereditary properties are examples of such properties; i.e., they are close to being semilinear (Lemma 31). Combining Theorem 25 and Lemma 31, we get an alternative proof that hyperfinite hereditary properties are uniformly testable in constant time in the $\mathrm{BDRD}_{+/-}$ model (Theorem 28). We believe that this highlights that semilinearity of neighborhood histograms is a natural and powerful concept. In the sketch of the proof that every hyperfinite hereditary property is constant time testable in the BDRD model (which uses methods similar to [7] and [12]), we start by testing for hyperfiniteness, which is done by extending the methods of [7]. In the tester for hyperfiniteness, an estimate of the input database's neighborhood distribution vector is computed, which is then compared against a δ-net of all the neighborhood distribution vectors of hyperfinite databases (where δ is some small constant). If the input is declared to be hyperfinite, we then sample a constant number of elements and check if the induced sub-database on the union of the (fixed radius) neighborhoods of the sampled elements is in the property. In our alternative tester, an estimate of the neighborhood distribution vector of the input database is computed and then compared against the neighborhood distributions of constant size databases in the property.

Alon [24, Proposition 19.10] proved that for every bounded degree graph $\mathcal{G}$ there exists a constant size graph $\mathcal{H}$ that has a similar neighborhood distribution to $\mathcal{G}$. However, the proof is based on a compactness argument and does not give an explicit upper bound on the size of $\mathcal{H}$. Finding such a bound was suggested by Alon as an open problem [20]. We ask under which conditions on a given property **P** for every member of **P** there exists a constant size database with a similar neighborhood distribution that is also in **P**. It is known that hyperfinite properties are *local* in the BDRD model [1, 25], where a property is local if for any given database $\mathcal{D}$ from the input class, if $\mathcal{D}$ has a similar r-histogram to some database in the property, then $\mathcal{D}$ must be close to the property. We define a similar notion for the $\mathrm{BDRD}_{+/-}$ model and show that hyperfinite properties are also local in the $\mathrm{BDRD}_{+/-}$ model (Theorem 11). Using the locality of hyperfinite properties, we show that for any property **P** that is hyperfinite on the input class **C** and whose r-histograms are semilinear, if a database $\mathcal{D}$ is in **P**, then there exists a constant size database $\mathcal{D}'$ in **P** with a similar neighborhood distribution, but this is not true for databases in **C** that are far from **P**. Furthermore, we obtain upper and lower bounds on the size of $\mathcal{D}'$. We can then use this result to construct constant time testers. We first use the algorithm $\mathrm{EstimateFrequencies}_{r,s}$ (given in [25] and adapted to databases in [1]) to approximate the neighborhood distribution of the input database. Then we only have to check if the estimated distribution is close to the neighborhood distribution of a constant size database in the property.

As a corollary (Corollary 17), we obtain an explicit bound on the size on graphs $\mathcal{H}$ from Alon's theorem for "semilinear" properties, i.e., properties, where the histogram vectors of the neighborhood distributions form a semilinear set.

Further related work. Other than the work already mentioned in [1], there are only a handful of results on relational databases that utilize models from property testing. Chen and Yoshida [9] study a model that is close to the general graph model (cf., e.g., [3]) in which they study the testability of homomorphism inadmissibility. Ben-Moshe et al. [6] study the testability of near-sortedness (a property of relations that states that most tuples are close to their place in some desired order). Our model differs from both of these, as it relies on a degree bound and uses different types of oracle access. Explicit bounds for Alon's theorem restricted to high-girth graphs were given in [14].

Obtaining a characterization of constant query testable properties is a long-standing open problem. Ito et al. [21] give a characterization of the one-sided error constant query testable monotone and hereditary graph properties in the bounded degree (directed and undirected) graph model. Fichtenberger et al. [15] show that every constant query testable property in the bounded degree graph model either is finite or contains an infinite hyperfinite subproperty.

Organization. In Section 2 we introduce relevant notions used throughout the article. In Section 3 we introduce our property testing model for relational databases of bounded degree and we compare it to the classical model. In Section 4 we define the notion of locality in both the BDRD and BDRD$_{+/-}$ models and prove that hyperfinite properties are local in the BDRD$_{+/-}$ model. In Section 5 we prove our main theorems. In Section 6 we prove a more general theorem of the theorem proved in Section 5. Finally, in Section 7 we give an alternative proof of the constant time uniform testability of hyperfinite hereditary properties in the BDRD$_{+/-}$ model.

2 PRELIMINARIES

We let $\mathbb{N}$ be the set of natural numbers including 0, and $\mathbb{N}_{\geq 1} = \mathbb{N} \setminus \{0\}$. For each $n \in \mathbb{N}_{\geq 1}$, we let $[n] = \{1, 2, \ldots, n\}$.

Databases. A *schema* is a finite set $\sigma = \{R_1, \ldots, R_{|\sigma|}\}$ of relation names, where each $R \in \sigma$ has an *arity* $\mathrm{ar}(R) \in \mathbb{N}_{\geq 1}$. A *database* $\mathcal{D}$ of schema σ (σ-db for short) is of the form $\mathcal{D} = (D, R_1^{\mathcal{D}}, \ldots, R_{|\sigma|}^{\mathcal{D}})$, where D is a finite set, the set of *elements* of $\mathcal{D}$, and $R_i^{\mathcal{D}}$ is an $\mathrm{ar}(R_i)$-ary relation on D. The set D is also called the *domain* of $\mathcal{D}$. An *(undirected) graph* $\mathcal{G}$ is a tuple $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$, where $V(\mathcal{G})$ is a set of *vertices* and $E(\mathcal{G})$ is a set of two-element subsets of $V(\mathcal{G})$ (the *edges* of $\mathcal{G}$). An undirected graph can be seen as an $\{E\}$-db, where E is a binary relation name, interpreted by a symmetric, irreflexive relation.

We assume that all databases are linearly ordered or, equivalently, that $D = [n]$ for some $n \in \mathbb{N}$ (similar to [22]). We extend this linear ordering to a linear order on the relations of $\mathcal{D}$ via lexicographic ordering. The *Gaifman graph* of a σ-db $\mathcal{D}$ is the undirected graph $\mathcal{G}(\mathcal{D}) = (V, E)$, with vertex set $V := D$ and an edge between vertices a and b whenever $a \neq b$ and there is an $R \in \sigma$ and a tuple $(a_1, \ldots, a_{\mathrm{ar}(R)}) \in R^{\mathcal{D}}$ with $a, b \in \{a_1, \ldots, a_{\mathrm{ar}(R)}\}$. The *degree* $\deg(a)$ of an element a in a database $\mathcal{D}$ is the total number of tuples in all relations of $\mathcal{D}$ that contain a. We say the *degree* $\deg(\mathcal{D})$ of a database $\mathcal{D}$ is the maximum degree of its elements. A class of databases $\mathbf{C}$ has *bounded degree* if there exists a constant $d \in \mathbb{N}$ such that for all $\mathcal{D} \in \mathbf{C}$, $\deg(\mathcal{D}) \leq d$. (We always assume that classes of databases are closed under isomorphism.) Let us remark that the $\deg(\mathcal{D})$ and the (graph-theoretic) degree of $\mathcal{G}(\mathcal{D})$ only differ by at most a constant factor (cf., e.g., [13]). Since we only consider fixed finite schemas, both measures yield the same classes of relational structures of bounded degree (cf. the discussion in [13]). We define the *tree-width* of a database $\mathcal{D}$ as the tree-width of its Gaifman graph. (See, e.g., [17] for a discussion of tree-width in this context.) A class $\mathbf{C}$ of databases has *bounded tree-width* if there exists a constant $t \in \mathbb{N}$ such that all databases $\mathcal{D} \in \mathbf{C}$ have tree-width at most t. Let $\mathcal{D}$ be a σ-db, and $M \subseteq D$. The sub-database of $\mathcal{D}$ *induced by* M is the database $\mathcal{D}[M]$ with domain M and $R^{\mathcal{D}[M]} := R^{\mathcal{D}} \cap M^{\text{ar}(R)}$ for every $R \in \sigma$. An (ϵ, k)-*partition* of a σ-db $\mathcal{D}$ on n elements is a σ-db $\mathcal{D}'$ formed by removing at most ϵn many tuples from $\mathcal{D}$ such that every connected component in $\mathcal{D}'$ contains at most k elements. A class of σ-dbs $\mathbf{C} \subseteq \mathbf{D}$ is ρ-*hyperfinite* on $\mathbf{D}$ if for every $\epsilon \in (0, 1]$ and $\mathcal{D} \in \mathbf{C}$ there exists an $(\epsilon, \rho(\epsilon))$-partition $\mathcal{D}' \in \mathbf{D}$ of $\mathcal{D}$. We call $\mathbf{C}$ *hyperfinite* on $\mathbf{D}$ if there exists a function ρ such that $\mathbf{C}$ is ρ-hyperfinite on $\mathbf{D}$.

Logics. We shall only briefly introduce FO and CMSO. Detailed introductions can be found in [23] and [11]. Let **var** be a countable infinite set of *variables*, and fix a relational schema σ. The set FO$[\sigma]$ is built from *atomic formulas* of the form $x_1 = x_2$ or $R(x_1, \ldots, x_{\text{ar}(R)})$, where $R \in \sigma$ and $x_1, \ldots, x_{\text{ar}(R)} \in \mathbf{var}$, and is closed under Boolean connectives $(\neg, \vee, \wedge, \rightarrow, \leftrightarrow)$ and existential and universal quantifications $(\exists, \forall)$. ***Monadic second-order logic* (MSO)** is the extension of first-order logic that also allows quantification over subsets of the domain. CMSO extends MSO by allowing first-order modular counting quantifiers $\exists^m$ for every integer m (where $\exists^m \phi$ is true in a σ-db if the number of its elements for which ϕ is satisfied is divisible by m). A *free variable* of a formula is an (individual or set) variable that does not appear in the scope of a quantifier. A formula without free variables is called a *sentence*. For a σ-db $\mathcal{D}$ and a sentence ϕ we write $\mathcal{D} \models \phi$ to denote that $\mathcal{D}$ satisfies ϕ.

Proviso. *For the rest of the article, we fix a schema σ and numbers $d, t \in \mathbb{N}$ with $d \geq 2$. From now on, all databases are σ-dbs and have degree at most d, unless stated otherwise. We use $\mathbf{C}_d$ to denote the class of* all *σ-dbs with degree at most d, $\mathbf{C}_d^t$ to denote the class of* all *σ-dbs with degree at most d and tree-width at most t and finally we use $\mathbf{C}$ to denote a class of σ-dbs with degree at most d (i.e., $\mathbf{C}$ is any subset of $\mathbf{C}_d$).*

Property testing. Adler and Harwath [1] introduced the model of property testing for bounded degree relational databases, which is a straightforward extension of the model for bounded degree graphs [18]. We call this model the BDRD *model* for short, which we shall discuss below.

Property testing algorithms do not have access to the whole input database. Instead, they are given access via an *oracle*. Let $\mathcal{D}$ be an input σ-db on n elements. A property testing algorithm receives the number n as input, and it can make *oracle queries*[1] of the form (R, i, j), where $R \in \sigma$, $i \leq n$, and $j \leq \deg(\mathcal{D})$. The answer to (R, i, j) is the j^{th} tuple in $R^{\mathcal{D}}$ containing the i^{th} element[2] of $\mathcal{D}$ (if such a tuple does not exist then it returns $\perp$). We assume oracle queries are answered in constant time.

Let $\mathcal{D}, \mathcal{D}'$ be two σ-dbs, both having n elements. In the BDRD model the *distance* between $\mathcal{D}$ and $\mathcal{D}'$, denoted by $\text{dist}(\mathcal{D}, \mathcal{D}')$, is the minimum number of tuples that have to be inserted or removed from relations of $\mathcal{D}$ and $\mathcal{D}'$ to make $\mathcal{D}$ and $\mathcal{D}'$ isomorphic. For $\epsilon \in [0, 1]$, we say $\mathcal{D}$ and $\mathcal{D}'$ are ϵ-*close* if $\text{dist}(\mathcal{D}, \mathcal{D}') \leq \epsilon d n$, and $\mathcal{D}$ and $\mathcal{D}'$ are ϵ-*far* otherwise. A *property* is simply an isomorphism-closed class of databases. Note that every CMSO sentence ϕ defines a property $\mathbf{P}_\phi = \{\mathcal{D} \mid \mathcal{D} \models \phi\}$. We call $\mathbf{P}_\phi \cap \mathbf{C}$ the property *defined by ϕ on* $\mathbf{C}$. A σ-db $\mathcal{D}$ is ϵ-*close* to a property $\mathbf{P}$ if there exists a database $\mathcal{D}' \in \mathbf{P}$ that is ϵ-close to $\mathcal{D}$; otherwise $\mathcal{D}$ is ϵ-*far* from $\mathbf{P}$.

Let $\mathbf{P} \subseteq \mathbf{C}$ be a property and $\epsilon \in (0, 1]$ be the proximity parameter. An ϵ-*tester* for $\mathbf{P}$ on $\mathbf{C}$ is a probabilistic algorithm that is given oracle access to a σ-db $\mathcal{D} \in \mathbf{C}$ and it is given $n := |D|$ as auxiliary input. The algorithm does the following:

[1] Note that an oracle query is not a database query.
[2] According to the assumed linear order on D.

(1) If $\mathcal{D} \in \mathbf{P}$, then the tester accepts with probability at least 2/3.
(2) If $\mathcal{D}$ is ϵ-far from $\mathbf{P}$, then the tester rejects with probability at least 2/3.

The *query complexity* of a tester is the maximum number of oracle queries made. A tester has *constant* query complexity if the query complexity does not depend on the size of the input database. We say a property $\mathbf{P} \subseteq \mathbf{C}$ is *uniformly testable* in time $f(n)$ on $\mathbf{C}$ if for every $\epsilon \in (0, 1]$ there exists an ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ that has constant query complexity and whose running time on databases on n elements is $f(n)$. Note that this tester must work for all n.

Neighborhoods. For a σ-db $\mathcal{D}$ and $a, b \in D$, the *distance* between a and b in $\mathcal{D}$, denoted by $\mathrm{dist}_{\mathcal{D}}(a, b)$, is the length of a shortest path between a and b in $\mathcal{G}(\mathcal{D})$. Let $r \in \mathbb{N}$. For an element $a \in D$, we let $N_r^{\mathcal{D}}(a)$ denote the set of all elements of $\mathcal{D}$ that are at distance at most r from a. The *r-neighborhood* of a in $\mathcal{D}$, denoted by $\mathcal{N}_r^{\mathcal{D}}(a)$, is the tuple $(\mathcal{D}[N_r(a)], a)$, where a is called the *center*. We omit the superscript and write $N_r(a)$ and $\mathcal{N}_r(a)$ if $\mathcal{D}$ is clear from the context. Two r-neighborhoods, $\mathcal{N}_r(a)$ and $\mathcal{N}_r(b)$, are *isomorphic* (written $\mathcal{N}_r(a) \cong \mathcal{N}_r(b)$) if there is an isomorphism between $\mathcal{D}[N_r(a)]$ and $\mathcal{D}[N_r(b)]$ that maps a to b. An $\cong$-equivalence-class of r-neighborhoods is called an *r-neighborhood type* (or *r-type* for short). We let $T_r^{\sigma,d}$ denote the set of all r-types with degree at most d, over schema σ. Note that for fixed d and σ, the cardinality $|T_r^{\sigma,d}| =: \mathrm{c}(r)$ is a constant, only depending on r and d. We say that an element $a \in D$ *has r-type* τ if $\mathcal{N}_r^{\mathcal{D}}(a) \in \tau$. For $r \in \mathbb{N}$, the *r-histogram* of a database $\mathcal{D}$, denoted by $\mathrm{h}_r(\mathcal{D})$, is the vector with $\mathrm{c}(r)$ components, indexed by the r-types, where the component corresponding to type τ contains the number of elements of $\mathcal{D}$ of r-type τ. The *r-neighborhood distribution* of $\mathcal{D}$, denoted by $\mathrm{dv}_r(\mathcal{D})$, is the vector $\mathrm{h}_r(\mathcal{D})/n$, where $|D| = n$. For a class of σ-dbs $\mathbf{C}$ and $r \in \mathbb{N}$, we let $\mathrm{h}_r(\mathbf{C}) := \{\mathrm{h}_r(\mathcal{D}) \mid \mathcal{D} \in \mathbf{C}\}$. A set is *semilinear* if it is a finite union of linear sets. A set $M \subseteq \mathbb{N}^c$ is linear if $M = \{\bar{v}_0 + a_1\bar{v}_1 + \cdots + a_k\bar{v}_k \mid a_1, \ldots, a_k \in \mathbb{N}\}$, for some $\bar{v}_0, \ldots, \bar{v}_k \in \mathbb{N}^c$. From a result in [16] about many-sorted spectra of CMSO sentences it can be derived that the set of r-histograms of properties defined by a CMSO sentence on $\mathbf{C}_d^t$ are semilinear.

Lemma 1 ([1, 16]). *For each $r \in \mathbb{N}$ and each property $\mathbf{P} \subseteq \mathbf{C}_d^t$ definable by a CMSO sentence on $\mathbf{C}_d^t$, the set $\mathrm{h}_r(\mathbf{P})$ is semilinear.*

Model of computation. We use **Random Access Machines (RAMs)** and a uniform cost measure when analyzing our algorithms; i.e., we assume all basic arithmetic operations including random sampling can be done in constant time, regardless of the size of the numbers involved.

3 THE MODEL

We shall now introduce our property testing model for bounded degree relational databases, which is an extension of the BDRD model discussed in Section 2. The notions of oracle queries, properties, ϵ-tester, query complexity, and uniform testability remain the same, but we have an alternative definition of distance and ϵ-closeness. In our model, which we shall call the $\mathrm{BDRD}_{+/-}$ *model* for short, we can add and remove elements as well as tuples and can therefore compare databases that are on a different number of elements.

Definition 2 (Distance and ϵ-closeness). Let $\mathcal{D}, \mathcal{D}' \in \mathbf{C}_d$ and $\epsilon \in [0, 1]$. The distance between $\mathcal{D}$ and $\mathcal{D}'$ (denoted by $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}')$) is the minimum number of modifications we need to make to $\mathcal{D}$ and $\mathcal{D}'$ to make them isomorphic where a modification is either (1) inserting a new element, (2) deleting an element (and as a result deleting any tuple that contains that element), (3) inserting a tuple, or (4) deleting a tuple. We then say $\mathcal{D}$ and $\mathcal{D}'$ are ϵ-close if $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \epsilon d \min\{|D|, |D'|\}$ and are ϵ-far otherwise.

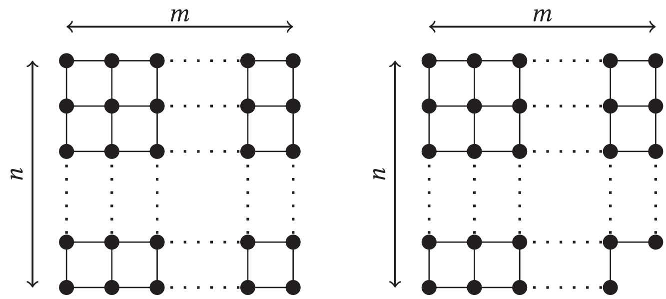

Fig. 1. The graphs $\mathcal{G}_{n,m}$ and $\mathcal{H}_{n,m}$ (respectively) of Example 3.

The following example illustrates the difference between the distance measure of the BDRD and the distance measure of the $\mathrm{BDRD}_{+/-}$ model.

Example 3. Let $\mathbf{P} = \{\mathcal{G}_{n,m} \mid n, m \in \mathbb{N}_{>1}\}$, where $\mathcal{G}_{n,m}$ is an n by m grid graph as shown in Figure 1. Let us consider the graph $\mathcal{H}_{n,m}$ for some $n, m \in \mathbb{N}$ that is formed from $\mathcal{G}_{n,m}$ by removing a corner vertex. In the $\mathrm{BDRD}_{+/-}$ model the distance between $\mathcal{H}_{n,m}$ and $\mathcal{G}_{n,m}$ is 1 (we remove a corner vertex from $\mathcal{G}_{n,m}$ to get $\mathcal{H}_{n,m}$) and therefore $\mathcal{H}_{n,m}$ is at distance 1 from $\mathbf{P}$ in the $\mathrm{BDRD}_{+/-}$ model. In the BDRD model, if two graphs are on a different number of vertices, then the distance between them is infinity. Therefore, if $nm - 1$ is a prime number, then $\mathcal{H}_{n,m}$ is at distance infinity from $\mathbf{P}$ in the BDRD model.

We now show that if a property is testable in the BDRD model, then it is also testable in the $\mathrm{BDRD}_{+/-}$ model, but the converse is not true. This allows for more testable properties in the $\mathrm{BDRD}_{+/-}$ model.

Lemma 4. *Let $\mathbf{P} \subseteq \mathbf{C}$. If $\mathbf{P}$ is uniformly testable on $\mathbf{C}$ in time $f(n)$ in the* BDRD *model, then $\mathbf{P}$ is also uniformly testable on $\mathbf{C}$ in time $f(n)$ in the* $\mathrm{BDRD}_{+/-}$ *model.*

Proof. Let π be an ϵ-tester, that runs in time $f(n)$, for $\mathbf{P}$ on $\mathbf{C}$ in the BDRD model. We claim that π is also an ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model. Let $\mathcal{D} \in \mathbf{C}$ be the input σ-db. If $\mathcal{D} \in \mathbf{P}$, then π will accept with probability at least 2/3. If $\mathcal{D}$ is ϵ-far from $\mathbf{P}$ in the $\mathrm{BDRD}_{+/-}$ model, then it must also be ϵ-far from $\mathbf{P}$ in the BDRD model and therefore π will reject with probability at least 2/3. Hence, π is an ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model. □

Theorem 5 ([18]). *In the bounded degree model, bipartiteness cannot be tested with query complexity $o(\sqrt{n})$, where n is the number of vertices of the input graph.*

Lemma 6. *There exists a class $\mathbf{C}$ of σ-dbs and a property $\mathbf{P} \subseteq \mathbf{C}$ such that $\mathbf{P}$ is trivially testable on $\mathbf{C}$ in the* $\mathrm{BDRD}_{+/-}$ *model but is not testable on $\mathbf{C}$ in the* BDRD *model.*

Proof. Let $\mathbf{C}$ be the class of all graphs with degree at most d. Let $\mathbf{P} = \mathbf{P}_1 \cup \mathbf{P}_2 \subseteq \mathbf{C}$ be the property where $\mathbf{P}_1$ contains all bipartite graphs in $\mathbf{C}$ and $\mathbf{P}_2$ contains all graphs in $\mathbf{C}$ that have an odd number of vertices. In the $\mathrm{BDRD}_{+/-}$ model every $\mathcal{G} \in \mathbf{C}$ is ϵ-close to $\mathbf{P}$ if $|V(\mathcal{G})| \geq 1/(\epsilon d)$ and hence $\mathbf{P}$ is trivially testable on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model (the tester accepts if $|V(\mathcal{G})| \geq 1/(\epsilon d)$ and does a full check of the input otherwise). In the BDRD model, if the input graph has an even number of vertices, then it is far from $\mathbf{P}_2$ and so we have to test for $\mathbf{P}_1$. By Theorem 5, bipartiteness is not testable (with constant query complexity) in the BDRD model. In particular, in the proof of Theorem 5, Goldreich and Ron show that for any even n there exists two families, $\mathcal{G}_1 \subseteq \mathbf{C}$ and $\mathcal{G}_2 \subseteq \mathbf{C}$, of n-vertex graphs such that every graph in $\mathcal{G}_1$ is bipartite and almost all graphs in $\mathcal{G}_2$ are far from being bipartite but any algorithm that performs $o(\sqrt{n})$ queries cannot distinguish between a graph chosen randomly from $\mathcal{G}_1$ and a graph chosen randomly from $\mathcal{G}_2$. Therefore, $\mathbf{P}$ is not testable on $\mathbf{C}$ in the BDRD model. □

Note that the underlying general principle of the above proof can be applied to obtain further examples of properties that are testable in the BDRD$_{+/-}$ model but not testable in the BDRD model.

4 LOCALITY OF PROPERTIES

It is known that every hyperfinite property is "local" in the BDRD model (Theorem 8), where a property is "local" if, *whenever* a σ-db $\mathcal{D}$ has *an* r-histogram *similar* to some σ-db (with the same domain size) that has the property, then $\mathcal{D}$ must be ϵ-close to the property [1, 25]. This is summarized in Definition 7 and Theorem 8 below. We define an equivalent definition of locality in the BDRD$_{+/-}$ model (Definition 9). We prove that any property that is local in the BDRD model is also local in the BDRD$_{+/-}$ model (Lemma 10) and hence every hyperfinite property is local in the BDRD$_{+/-}$ model (Theorem 11). Theorem 11 is essential for the proof of Theorem 20.

Definition 7 (Locality in the BDRD *Model).* Let $\epsilon \in (0, 1]$. A property $\mathbf{P} \subseteq \mathbf{C}$ is ϵ-*local* on $\mathbf{C}$ in the BDRD model if there exists $\lambda := \lambda_7(\epsilon) \in (0, 1]$, $r := r_7(\epsilon) \in \mathbb{N}$, and $N := N_7(\epsilon) \in \mathbb{N}$ such that for each $\mathcal{D} \in \mathbf{P}$ and $\mathcal{D}' \in \mathbf{C}$ with the same number $n \geq N$ of elements, if $\|h_r(\mathcal{D}) - h_r(\mathcal{D}')\|_1 \leq \lambda n$, then $\mathcal{D}'$ is ϵ-close to $\mathbf{P}$ in the BDRD model.

We call the parameters r and λ the *locality radius* and *disc proximity* of $\mathbf{P}$ for ϵ, respectively. A property is *local* in the BDRD model if it is ϵ-local in the BDRD model for every $\epsilon \in (0, 1]$.

THEOREM 8 ([1, 25]). *Let* $\mathbf{C}$ *be closed under removing tuples. If a property* $\mathbf{P} \subseteq \mathbf{C}$ *is hyperfinite on* $\mathbf{C}$*, then* $\mathbf{P}$ *is local on* $\mathbf{C}$ *in the* BDRD *model.*

We now define the notion of locality in the BDRD$_{+/-}$ model.

Definition 9 (Locality in the BDRD$_{+/-}$ *Model).* Let $\epsilon \in (0, 1]$. A property $\mathbf{P} \subseteq \mathbf{C}$ is ϵ-*local* on $\mathbf{C}$ in the BDRD$_{+/-}$ model if there exists $\lambda := \lambda_9(\epsilon) \in (0, 1]$, $r := r_9(\epsilon) \in \mathbb{N}$ and $N := N_9(\epsilon) \in \mathbb{N}$ such that for each $\mathcal{D} \in \mathbf{P}$ and $\mathcal{D}' \in \mathbf{C}$, on $|D| \geq N$ and $|D'| \geq N$ elements respectively, if $\|\mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq \lambda \min\{|D|, |D'|\}$, then $\mathcal{D}'$ is ϵ-close to $\mathbf{P}$ in the BDRD$_{+/-}$ model.

We call the parameters r and λ the *locality radius* and *disc proximity* of $\mathbf{P}$ for ϵ, respectively. A property is *local* in the BDRD$_{+/-}$ model if it is ϵ-local in the BDRD$_{+/-}$ model for every $\epsilon \in (0, 1]$.

LEMMA 10. *If a property* $\mathbf{P} \subseteq \mathbf{C}$ *is local on* $\mathbf{C}$ *in the* BDRD *model, then* $\mathbf{P}$ *is local on* $\mathbf{C}$ *in the* BDRD$_{+/-}$ *model.*

PROOF. Let $\epsilon \in (0, 1]$. Let $r_7(\epsilon/4)$, $\lambda_7(\epsilon/4)$, and $N_7(\epsilon/4)$ be as in Definition 7 for $\mathbf{P}$ and $\epsilon/4$. Let $r := r_7(\epsilon/4)$, let $N := N_7(\epsilon/4)$, and let

$$\lambda := \frac{\epsilon\lambda_7(\epsilon/4)}{1 + d^{r+1}}.$$

We will prove that $\mathbf{P}$ is ϵ-local on $\mathbf{C}$ in the BDRD$_{+/-}$ model with $r_9(\epsilon) = r$, $\lambda_9(\epsilon) = \lambda$, and $N_9(\epsilon) = N$.

Let $\mathcal{D} \in \mathbf{P}$ and $\mathcal{D}' \in \mathbf{C}$, where $|D| \geq N$ and $|D'| \geq N$. Let us assume that $\|\mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq \lambda \min\{|D|, |D'|\}$ and $\mathbf{P}$ is local on $\mathbf{C}$ in the BDRD model. We will show that $\mathcal{D}'$ is ϵ-close to $\mathbf{P}$.

If $|D| = |D'|$, then since $\lambda \leq \lambda_7(\epsilon/4)$, $r = r_7(\epsilon/4)$, and $N = N_7(\epsilon/4)$, $\mathcal{D}'$ is $\epsilon/4$-close to $\mathbf{P}$ and hence $\mathcal{D}'$ is also ϵ-close to $\mathbf{P}$. So let us assume that $|D| \neq |D'|$. Let $\mathcal{D}_1$ be the σ-db on $|D|$ elements formed from $\mathcal{D}'$ by either removing $|D'| - |D|$ elements if $|D| < |D'|$ or adding $|D| - |D'|$ new elements if $|D'| < |D|$. Note that as $\|\mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq \lambda \min\{|D|, |D'|\}$ and by definition $\|\mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 = \sum_{i=1}^{c(r)} |\mathrm{h}_r(\mathcal{D})[i] - \mathrm{h}_r(\mathcal{D}')[i]|$, we have $||D| - |D'|| \leq \lambda \min\{|D|, |D'|\}$. When an element a is removed, the r-type of any element in $N_r(a)$ will change. Since

$|N_r(a)| \leq d^{r+1}$ (cf., e.g., Lemma 3.2 (a) of [8]) and $||D| - |D'|| \leq \lambda \min\{|D|, |D'|\}$, we have $\| \mathrm{h}_r(\mathcal{D}') - \mathrm{h}_r(\mathcal{D}_1)\|_1 \leq \lambda \min\{|D|, |D'|\} d^{r+1}$. Therefore,

$$\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}_1)\|_1 \leq \lambda \min\{|D|, |D'|\}(1 + d^{r+1}) \leq \lambda_7(\epsilon/4)|D|$$

by the choice of λ. Since $\mathbf{P}$ is local on $\mathbf{C}$ in the BDRD model, $|D| = |D_1|$, and $\mathcal{D} \in \mathbf{P}$, $\mathcal{D}_1$ is $\epsilon/4$-close to $\mathbf{P}$ in the BDRD model. Hence, there exists a σ-db $\mathcal{D}_2 \in \mathbf{P}$ such that $|D_2| = |D|$ and $\mathrm{dist}(\mathcal{D}_1, \mathcal{D}_2) \leq \epsilon d|D|/4$. By the definition of the two distance measures dist and $\mathrm{dist}_{+/-}$, we have $\mathrm{dist}_{+/-}(\mathcal{D}_1, \mathcal{D}_2) \leq \mathrm{dist}(\mathcal{D}_1, \mathcal{D}_2) \leq \epsilon d|D|/4$ and by the construction of $\mathcal{D}_1$ we have $\mathrm{dist}_{+/-}(\mathcal{D}', \mathcal{D}_1) \leq \lambda \min\{|D|, |D'|\} = \lambda \min\{|D'|, |D_2|\}$. Therefore,

$$\mathrm{dist}_{+/-}(\mathcal{D}', \mathcal{D}_2) \leq \frac{\epsilon d|D|}{4} + \lambda \min\{|D'|, |D_2|\} \leq \epsilon d \min\{|D'|, |D_2|\},$$

since $|D| \leq (1 + \lambda) \min\{|D|, |D'|\} \leq 2 \min\{|D|, |D'|\} = 2 \min\{|D'|, |D_2|\}$ (if $|D| < |D'|$, then clearly this holds; otherwise since $||D| - |D'|| \leq \lambda \min\{|D|, |D'|\}$, $|D| \leq |D'| + \lambda \min\{|D|, |D'|\} = (1 + \lambda) \min\{|D|, |D'|\})$ and $\lambda \leq \epsilon d/2$. Hence, in the $\mathrm{BDRD}_{+/-}$ model $\mathcal{D}'$ is ϵ-close to $\mathbf{P}$ as required. □

By combining Theorem 8 and Lemma 10 we obtain the following theorem.

THEOREM 11. *Let* $\mathbf{C}$ *be closed under removing tuples. If a property* $\mathbf{P} \subseteq \mathbf{C}$ *is hyperfinite on* $\mathbf{C}$*, then* $\mathbf{P}$ *is local on* $\mathbf{C}$ *in the* $\mathrm{BDRD}_{+/-}$ *model.*

5 MAIN RESULTS

We begin this section with the first of our main theorems (Theorem 12). We show that for any property $\mathbf{P}$ that is ϵ-local (in the $\mathrm{BDRD}_{+/-}$ model) on the input class $\mathbf{C}$, if the set of r-histograms of $\mathbf{P}$ is semilinear, then for every σ-db $\mathcal{D}$ in $\mathbf{P}$ there exists a constant size σ-db in $\mathbf{P}$ with a neighborhood distribution similar to that of $\mathcal{D}$, but this is not true for σ-dbs in $\mathbf{C}$ that are far from $\mathbf{P}$. We then use this result to prove that for such properties there exist ϵ-testers in the $\mathrm{BDRD}_{+/-}$ model that run in constant time (Theorem 19). As corollaries we obtain that hyperfinite properties whose set of r-histograms is semilinear are constant time testable (Theorem 20) and CMSO definable properties on σ-dbs of bounded tree-width and bounded degree are testable in constant time (Theorem 21).

THEOREM 12. *Let* $\epsilon \in (0, 1]$. *Let* $\mathbf{P} \subseteq \mathbf{C}$ *be a property that is* ϵ*-local on* $\mathbf{C}$ *(in the* $\mathrm{BDRD}_{+/-}$ *model) such that the set* $\mathrm{h}_r(\mathbf{P})$ *is semilinear, where* $r := r_9(\epsilon)$ *is the locality radius of* $\mathbf{P}$ *for* ϵ*. Then there exist* $n_{min} := n_{min}(\epsilon), n_{max} := n_{max}(\epsilon) \in \mathbb{N}$*, and* $f := f(\epsilon), \mu := \mu(\epsilon) \in (0, 1)$ *such that for every* $\mathcal{D} \in \mathbf{C}$ *with* $|D| > n_{max}$*,*

1. *if* $\mathcal{D} \in \mathbf{P}$*, then there exists a* $\mathcal{D}' \in \mathbf{P}$ *such that* $n_{min} \leq |D'| \leq n_{max}$ *and* $\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 \leq f - \mu$*, and*
2. *if* $\mathcal{D}$ *is* ϵ*-far from* $\mathbf{P}$ *(in the* $\mathrm{BDRD}_{+/-}$ *model), then for every* $\mathcal{D}' \in \mathbf{P}$ *such that* $n_{min} \leq |D'| \leq n_{max}$*, we have* $\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 > f + \mu$*.*

PROOF. Let $\lambda := \lambda_9(\epsilon)$ and $N := N_9(\epsilon)$ be as in Definition 9 for $\mathbf{P}$ and ϵ, and let $c := \mathrm{c}(r)$ (the number of r-types). First note that if $\mathbf{P}$ is empty, then for any choice of $n_{\min}$, $n_{\max}$, f, and μ, both 1 and 2 in the theorem statement are true and hence we shall assume that $\mathbf{P}$ is non-empty. As $\mathrm{h}_r(\mathbf{P})$ is a semilinear set, we can write it as follows: $\mathrm{h}_r(\mathbf{P}) = M_1 \cup M_2 \cup \cdots \cup M_m$, where $m \in \mathbb{N}$, and for each $i \in [m]$, $M_i = \{\bar{v}_0^i + a_1\bar{v}_1^i + \cdots + a_{k_i}\bar{v}_{k_i}^i \mid a_1, \ldots, a_{k_i} \in \mathbb{N}\}$ is a linear set, where $\bar{v}_0^i, \ldots, \bar{v}_{k_i}^i \in \mathbb{N}^c$, and for each $j \in [k_i]$, $\|\bar{v}_j^i\|_1 \neq 0$. Let $k := \max_{i \in [m]} k_i + 1$ and $v := \max_{i \in [m]}(\max_{j \in [0, k_i]} \|\bar{v}_j^i\|_1)$ (note that $v > 0$ as $\mathbf{P}$ is non-empty). Let $n_{\min} := n_0 - kv$, $n_{\max} := n_0 + kv$, $f := \frac{\lambda}{3c}$, and $\mu := \frac{\lambda}{6c}$, where

$$n_0 := \max\left\{\frac{9N}{5}, kv\left(\frac{3ckv}{f - \mu} + 1\right)\right\}.$$

Note that $n_{\min} > 0$ by the choice of n_0, f, and μ.

(Proof of 1.) Assume $\mathcal{D} \in \mathbf{P}$ and $|D| = n > n_{\max}$. Then there exists some $i \in [m]$ and $a_1^{\mathcal{D}}, \dots, a_{k_i}^{\mathcal{D}} \in \mathbb{N}$ such that $\mathrm{h}_r(\mathcal{D}) = \bar{v}_0^i + a_1^{\mathcal{D}}\bar{v}_1^i + \cdots + a_{k_i}^{\mathcal{D}}\bar{v}_{k_i}^i$ (note that $n = \|\bar{v}_0^i\|_1 + \sum_{j\in[k_i]} a_j^{\mathcal{D}}\|\bar{v}_j^i\|_1$). Let $\mathcal{D}'$ be the σ-db with r-histogram $\bar{v}_0^i + a_1^{\mathcal{D}'}\bar{v}_1^i + \cdots + a_{k_i}^{\mathcal{D}'}\bar{v}_{k_i}^i \in M_i$, where $a_j^{\mathcal{D}'}$ is the nearest integer to $a_j^{\mathcal{D}} n_0/n$, and hence $a_j^{\mathcal{D}} n_0/n - 1/2 \le a_j^{\mathcal{D}'} \le a_j^{\mathcal{D}} n_0/n + 1/2$. Note that since $\bar{v}_0^i + a_1^{\mathcal{D}'}\bar{v}_1^i + \cdots + a_{k_i}^{\mathcal{D}'}\bar{v}_{k_i}^i \in \mathrm{h}_r(\mathbf{P})$, $\mathcal{D}'$ exists and $\mathcal{D}' \in \mathbf{P}$. We need to show that $n_{\min} \le |D'| \le n_{\max}$ and $\|\,\mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 \le f - \mu$.

Claim 13. $|D'| \ge n_{min}$.

Proof. By the choice of $a_j^{\mathcal{D}'}$ for $j \in [k_i]$,

$$\begin{aligned}
|D'| &= \left\|\bar{v}_0^i\right\|_1 + \sum_{j\in[k_i]} a_j^{\mathcal{D}'}\left\|\bar{v}_j^i\right\|_1 \\
&\ge \left\|\bar{v}_0^i\right\|_1 + \sum_{j\in[k_i]} \left(\frac{a_j^{\mathcal{D}} n_0}{n} - \frac{1}{2}\right)\left\|\bar{v}_j^i\right\|_1 \\
&= \left\|\bar{v}_0^i\right\|_1 - \frac{1}{2}\sum_{j\in[k_i]}\left\|\bar{v}_j^i\right\|_1 + \frac{n_0}{n}\sum_{j\in[k_i]} a_j^{\mathcal{D}}\left\|\bar{v}_j^i\right\|_1 \\
&= \left\|\bar{v}_0^i\right\|_1 - \frac{1}{2}\sum_{j\in[k_i]}\left\|\bar{v}_j^i\right\|_1 + n_0 - \frac{n_0\left\|\bar{v}_0^i\right\|_1}{n} \\
&\ge \left\|\bar{v}_0^i\right\|_1 - \frac{1}{2}\sum_{j\in[k_i]}\left\|\bar{v}_j^i\right\|_1 + n_0 - \left\|\bar{v}_0^i\right\|_1 \\
&\ge -kv + n_0 = n_{\min},
\end{aligned}$$

as $\sum_{j\in[k_i]} a_j^{\mathcal{D}}\|\bar{v}_j^i\|_1 = n - \|\bar{v}_0^i\|_1$ and $n > n_{\max} \ge n_0$. □

Claim 14. $|D'| \le n_{max}$.

Proof. By the choice of $a_j^{\mathcal{D}'}$ for $j \in [k_i]$,

$$\begin{aligned}
|D'| &= \left\|\bar{v}_0^i\right\|_1 + \sum_{j\in[k_i]} a_j^{\mathcal{D}'}\left\|\bar{v}_j^i\right\|_1 \\
&\le \left\|\bar{v}_0^i\right\|_1 + \sum_{j\in[k_i]} \left(\frac{a_j^{\mathcal{D}} n_0}{n} + \frac{1}{2}\right)\left\|\bar{v}_j^i\right\|_1 \\
&= \left\|\bar{v}_0^i\right\|_1 + \frac{1}{2}\sum_{j\in[k_i]}\left\|\bar{v}_j^i\right\|_1 + n_0\left(1 - \frac{\left\|\bar{v}_0^i\right\|_1}{n}\right) \\
&\le \sum_{0\le j\le k_i}\left\|\bar{v}_j^i\right\|_1 + n_0 \\
&\le kv + n_0 = n_{\max},
\end{aligned}$$

as $\sum_{j\in[k_i]} a_j^{\mathcal{D}}\|\bar{v}_j^i\|_1 = n - \|\bar{v}_0^i\|_1$. □

Claim 15. $\|\,\mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 \le f - \mu$.

Proof. By definition, $\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}')\|_1 = \sum_{j\in[c]} |\operatorname{dv}_r(\mathcal{D})[j] - \operatorname{dv}_r(\mathcal{D}')[j]|$. First recall that $0 < n_0 - kv \le |D'| \le n_0 + kv < n$ and note that for every $\ell \in [k_i]$, $a_\ell^{\mathcal{D}} \le n$ (since $\|\bar{v}_\ell^i\|_1 \neq 0$). Then for every $j \in [c]$, by the choice of $a_\ell^{\mathcal{D}'}$ for $\ell \in [k_i]$,

$$\begin{aligned}
\operatorname{dv}_r(\mathcal{D})[j] - \operatorname{dv}_r(\mathcal{D}')[j] &= \bar{v}_0^i[j]\left(\frac{1}{n} - \frac{1}{|D'|}\right) + \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}}}{n} - \frac{a_\ell^{\mathcal{D}'}}{|D'|}\right)\\
&\le \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}}}{n} - \frac{a_\ell^{\mathcal{D}} n_0}{n|D'|} + \frac{1}{2|D'|}\right)\\
&= \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}}}{n}\left(\frac{|D'| - n_0}{|D'|}\right) + \frac{1}{2|D'|}\right)\\
&\le \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{n}{n}\left(\frac{kv + n_0 - n_0}{n_0 - kv}\right) + \frac{1}{2(n_0 - kv)}\right)\\
&= \left(\frac{2kv+1}{2(n_0 - kv)}\right)\sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\\
&\le \frac{kv(2kv+1)}{n_0 - kv}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\operatorname{dv}_r(\mathcal{D})[j] - \operatorname{dv}_r(\mathcal{D}')[j] &\ge -\frac{\bar{v}_0^i[j]}{|D'|} + \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}}}{n}\left(\frac{|D'| - n_0}{|D'|}\right) - \frac{1}{2|D'|}\right)\\
&\ge -\frac{\bar{v}_0^i[j]}{|D'|} + \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}}}{n}\left(\frac{-kv + n_0 - n_0}{|D'|}\right) - \frac{1}{2|D'|}\right)\\
&= -\frac{\bar{v}_0^i[j]}{|D'|} - \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}} kv}{n|D'|} + \frac{1}{2|D'|}\right)\\
&\ge -\frac{\bar{v}_0^i[j]}{n_0 - kv} - \sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\left(\frac{nkv}{n(n_0 - kv)} + \frac{1}{2(n_0 - kv)}\right)\\
&= -\frac{\bar{v}_0^i[j]}{n_0 - kv} - \left(\frac{2kv+1}{2(n_0 - kv)}\right)\sum_{\ell\in[k_i]} \bar{v}_\ell^i[j]\\
&\ge -\frac{kv(2kv+1)}{n_0 - kv}.
\end{aligned}$$

Hence,

$$|\operatorname{dv}_r(\mathcal{D})[j] - \operatorname{dv}_r(\mathcal{D}')[j]| \le \frac{kv(2kv+1)}{n_0 - kv} \le \frac{3(kv)^2}{n_0 - kv} \le \frac{f-\mu}{c}$$

by the choice of n_0. Therefore,

$$\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}')\|_1 = \sum_{j\in[c]} |\operatorname{dv}_r(\mathcal{D})[j] - \operatorname{dv}_r(\mathcal{D}')[j]| \le f - \mu,$$

as required. □

(Proof of 2.) Assume $\mathcal{D}$ is ϵ-far from $\mathbf{P}$ and $|D| = n > n_{\max}$. For a contradiction let us assume there does exist a σ-db $\mathcal{D}' \in \mathbf{P}$ such that $n_{\min} \leq |D'| \leq n_{\max}$ and $\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}')\|_1 \leq f + \mu$. As $\mathcal{D}' \in \mathbf{P}$ there exists some $i \in [m]$ and $a_1^{\mathcal{D}'}, \dots, a_{k_i}^{\mathcal{D}'} \in \mathbb{N}$ such that $\operatorname{h}_r(\mathcal{D}') = \bar{v}_0^i + a_1^{\mathcal{D}'}\bar{v}_1^i + \dots + a_{k_i}^{\mathcal{D}'}\bar{v}_{k_i}^i$. For every $j \in [k_i]$ let $a_j^{\mathcal{D}''}$ be the nearest integer to $a_j^{\mathcal{D}'} n/|D'|$. Then $\bar{v}_0^i + a_1^{\mathcal{D}''}\bar{v}_1^i + \dots + a_{k_i}^{\mathcal{D}''}\bar{v}_{k_i}^i \in \operatorname{h}_r(\mathbf{P})$, and hence there exists a σ-db in $\mathbf{P}$ with r-histogram $\bar{v}_0^i + a_1^{\mathcal{D}''}\bar{v}_1^i + \dots + a_{k_i}^{\mathcal{D}''}\bar{v}_{k_i}^i$. Let $\mathcal{D}''$ be one such σ-db.

Claim 16. *$\mathcal{D}$ is ϵ-close to* $\mathbf{P}$.

Proof. First note that as $\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}')\|_1 \leq f + \mu$ and $\operatorname{h}_r(\mathcal{D}') = \bar{v}_0^i + a_1^{\mathcal{D}'}\bar{v}_1^i + \dots + a_{k_i}^{\mathcal{D}'}\bar{v}_{k_i}^i$, for every $j \in [c]$

$$\frac{\bar{v}_0^i[j] + \sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}'}\bar{v}_\ell^i[j]}{|D'|} - f - \mu \leq \operatorname{dv}_r(\mathcal{D})[j] \leq \frac{\bar{v}_0^i[j] + \sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}'}\bar{v}_\ell^i[j]}{|D'|} + f + \mu,$$

and therefore,

$$n\left(\frac{\bar{v}_0^i[j] + \sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}'}\bar{v}_\ell^i[j]}{|D'|} - f - \mu\right) \leq \operatorname{h}_r(\mathcal{D})[j] \leq n\left(\frac{\bar{v}_0^i[j] + \sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}'}\bar{v}_\ell^i[j]}{|D'|} + f + \mu\right).$$

Hence, by the choice of $a_\ell^{\mathcal{D}''}$ for $\ell \in [k_i]$,

$$\begin{aligned}
\operatorname{h}_r(\mathcal{D})[j] - \operatorname{h}_r(\mathcal{D}'')[j] &\leq \bar{v}_0^i[j]\left(\frac{n}{|D'|} - 1\right) + \sum_{\ell \in [k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} - a_\ell^{\mathcal{D}''}\right) + fn + \mu n \\
&\leq \bar{v}_0^i[j]\frac{n}{|D'|} + \sum_{\ell \in [k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} - \left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} - \frac{1}{2}\right)\right) + fn + \mu n \\
&= \bar{v}_0^i[j]\frac{n}{|D'|} + \frac{1}{2}\sum_{\ell \in [k_i]} \bar{v}_\ell^i[j] + fn + \mu n.
\end{aligned}$$

Similarly, by the choice of $a_\ell^{\mathcal{D}''}$ for $\ell \in [k_i]$ and as $n > |D'|$,

$$\begin{aligned}
\operatorname{h}_r(\mathcal{D})[j] - \operatorname{h}_r(\mathcal{D}'')[j] &\geq \bar{v}_0^i[j]\left(\frac{n}{|D'|} - 1\right) + \sum_{\ell \in [k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} - a_\ell^{\mathcal{D}''}\right) - fn - \mu n \\
&\geq -\bar{v}_0^i[j]\frac{n}{|D'|} + \sum_{\ell \in [k_i]} \bar{v}_\ell^i[j]\left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} - \left(\frac{a_\ell^{\mathcal{D}'}n}{|D'|} + \frac{1}{2}\right)\right) - fn - \mu n \\
&= -\bar{v}_0^i[j]\frac{n}{|D'|} - \frac{1}{2}\sum_{\ell \in [k_i]} \bar{v}_\ell^i[j] - fn - \mu n.
\end{aligned}$$

Therefore,

$$\begin{aligned}
|\operatorname{h}_r(\mathcal{D})[j] - \operatorname{h}_r(\mathcal{D}'')[j]| &\leq \bar{v}_0^i[j]\frac{n}{|D'|} + \frac{1}{2}\sum_{\ell \in [k_i]} \bar{v}_\ell^i[j] + fn + \mu n \\
&\leq \frac{n}{|D'|}\sum_{0 \leq \ell \leq k_i} \bar{v}_\ell^i[j] + fn + \mu n \\
&\leq \frac{nkv}{|D'|} + fn + \mu n
\end{aligned}$$

$$
\begin{aligned}
&= n\left(\frac{kv}{|D'|} + \frac{\lambda}{3c} + \frac{\lambda}{6c}\right)\\
&\leq n\left(\frac{\lambda}{18c} + \frac{\lambda}{3c} + \frac{\lambda}{6c}\right)\\
&= \frac{5\lambda n}{9c}
\end{aligned}
$$

by the choice of f and μ and as

$$|D'| \geq n_{\min} \geq \frac{3c(kv)^2}{f-\mu} = \frac{18(ckv)^2}{\lambda} \geq \frac{18ckv}{\lambda}.$$

As $\mathbf{P}$ is ϵ-local on $\mathbf{C}$, if $|D''| \geq N$ and $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}'')\|_1 \leq \lambda \min\{n, |D''|\}$, then $\mathcal{D}$ is ϵ-close to $\mathbf{P}$. If $| \mathrm{h}_r(\mathcal{D})[j] - \mathrm{h}_r(\mathcal{D}'')[j]| \leq \frac{\lambda}{c}\min\{n, |D''|\}$, then $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}'')\|_1 \leq \lambda \min\{n, |D''|\}$. Clearly, $\frac{5\lambda n}{9c} < \frac{\lambda n}{c}$. We also have

$$
\begin{aligned}
|D''| &= \left\|\bar{v}_0^i\right\|_1 + \sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}''}\left\|\bar{v}_\ell^i\right\|_1\\
&\geq \left\|\bar{v}_0^i\right\|_1 + \sum_{\ell \in [k_i]} \left(\frac{a_\ell^{\mathcal{D}'} n}{|D'|} - \frac{1}{2}\right)\left\|\bar{v}_\ell^i\right\|_1\\
&= \left\|\bar{v}_0^i\right\|_1 - \frac{1}{2}\sum_{\ell \in [k_i]}\left\|\bar{v}_\ell^i\right\|_1 + \frac{n}{|D'|}\sum_{\ell \in [k_i]} a_\ell^{\mathcal{D}'}\left\|\bar{v}_\ell^i\right\|_1\\
&\geq -kv + \frac{n}{|D'|}\left(|D'| - \left\|\bar{v}_0^i\right\|_1\right)\\
&\geq -\frac{n}{18} + \frac{17}{18}n\\
&> \frac{5n}{9}
\end{aligned}
$$

as

$$|D'| \geq \frac{18ckv}{\lambda} \geq 18v \geq 18\|\bar{v}_0^i\|_1 \text{ and } kv \leq \frac{(ckv)^2}{\lambda} = \frac{n_{\min}}{18} \leq \frac{n}{18}.$$

Therefore, $\frac{5\lambda n}{9c} \leq \frac{\lambda |D''|}{c}$ and hence $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}'')\|_1 \leq \lambda \min\{n, |D''|\}$. Furthermore, by the choice of $n_{\max}$, $\frac{5n}{9} \geq N$ and hence $|D''| \geq N$. Therefore, $\mathcal{D}$ is ϵ-close to $\mathbf{P}$. □

Claim 16 gives us a contradiction and therefore for every $\mathcal{D}' \in \mathbf{P}$ such that $n_{\min} \leq |D'| \leq n_{\max}$, we have $\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 > f + \mu$ as required. □

As mentioned in the introduction, Alon [24, Proposition 19.10] proved that on bounded degree graphs, for any graph $\mathcal{G}$, radius r, and $\epsilon > 0$ there always exists a graph $\mathcal{H}$ whose size is independent of $|V(\mathcal{G})|$ and whose r-neighborhood distribution vector satisfies $\| \mathrm{dv}_r(\mathcal{G}) - \mathrm{dv}_r(\mathcal{H})\|_1 \leq \epsilon$. However, the proof is only existential and does not provide an explicit bound on the size of $\mathcal{H}$. As a corollary to the proof of Theorem 12, we immediately obtain explicit bounds for classes of graphs and relational databases of bounded degree whose histogram vectors form a semilinear set.

Corollary 17. *Let $\epsilon \in (0, 1]$, $r \in \mathbb{N}$, and $\mathcal{D}$ be a σ-db that belongs to a class of σ-dbs $\mathbf{C}$ such that the set $\mathrm{h}_r(\mathbf{C})$ is semilinear, i.e., $\mathrm{h}_r(\mathbf{C}) = M_1 \cup M_2 \cup \cdots \cup M_m$, where $m \in \mathbb{N}$, and for each $i \in [m]$,*

$M_i = \{\bar{v}_0^i + a_1\bar{v}_1^i + \cdots + a_{k_i}\bar{v}_{k_i}^i \mid a_1, \ldots, a_{k_i} \in \mathbb{N}\}$ is a linear set where $\bar{v}_0^i, \ldots, \bar{v}_{k_i}^i \in \mathbb{N}^{\mathrm{c}(r)}$. Then there exists a σ-db $\mathcal{D}_0$ such that

$$\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}_0)\|_1 \leq \epsilon \text{ and } |D_0| \leq kv\left(\frac{18c^2kv}{\epsilon} + 2\right),$$

where $c := \mathrm{c}(r)$, $k := \max_{i\in[m]} k_i + 1$, and $v := \max_{i\in[m]}(\max_{j\in[0,k_i]} \|\bar{v}_j^i\|_1)$.

Proof. The class $\mathbf{C}$ is ϵ-local on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model with $\lambda_9(\epsilon) = \epsilon$, $r_9(\epsilon) = r$, and $N_9(\epsilon) = 0$ (note that we can choose any values for $\lambda_9(\epsilon)$, $r_9(\epsilon)$, and $N_9(\epsilon)$ by definition). By Theorem 12 there exist $n_{\min} := n_{\min}(\epsilon), n_{\max} := n_{\max}(\epsilon) \in \mathbb{N}$, and $f := f(\epsilon), \mu := \mu(\epsilon) \in (0,1)$ such that if $|D| > n_{\max}$, there exists a $\mathcal{D}_0 \in \mathbf{C}$ such that $n_{\min} \leq |D_0| \leq n_{\max}$ and $\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}_0)\|_1 \leq f - \mu$. From the proof of Theorem 12, $n_{\max} := \max\{9N_9(\epsilon)/5, kv(3ckv/(f-\mu)+1)\} + kv = kv(3ckv/(f-\mu)+2)$, $f := \epsilon/3c$, and $\mu := \epsilon/6c$. Hence,

$$\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}_0)\|_1 \leq f - \mu = \frac{\epsilon}{6c} \leq \epsilon \text{ and } |D_0| \leq kv\left(\frac{3ckv}{f-\mu} + 2\right) = kv\left(\frac{18c^2kv}{\epsilon} + 2\right),$$

as required. If $|D| \leq n_{\max}$, then $\mathcal{D}_0 = \mathcal{D}$ satisfies the corollary statement. □

Our aim is to construct constant time testers for local properties whose set of r-histograms are semilinear. If we can approximate the r-neighborhood distribution of a σ-db, then by Theorem 12 we only need to check whether this distribution is close or not to the r-neighborhood distribution of some small constant size σ-db. We let $\mathrm{EstimateFrequencies}_{r,s}$ be the algorithm that, given oracle access to an input σ-db $\mathcal{D}$, samples s many elements uniformly and independently from D and computes their r-type. The algorithm then returns the r-neighborhood distribution vector of the sample.

Lemma 18 ([1]). *Let $\mathcal{D} \in \mathbf{C}_d$ be a σ-db on n elements, $\mu \in (0,1)$, and $r \in \mathbb{N}$. If $s \geq \mathrm{c}(r)^2/\mu^2 \cdot \ln(20\,\mathrm{c}(r))$, with probability at least $9/10$ the vector $\bar{v}$ returned by the algorithm $\mathrm{EstimateFrequencies}_{r,s}$ on input $\mathcal{D}$ satisfies $\|\bar{v} - \operatorname{dv}_r(\mathcal{D})\|_1 \leq \mu$.*

Theorem 19. *Let $\epsilon \in (0,1]$ and let $\mathbf{P} \subseteq \mathbf{C}$ be a property that is ϵ-local on $\mathbf{C}$ (in the $\mathrm{BDRD}_{+/-}$ model). If for each $r \in \mathbb{N}$ the set $\mathrm{h}_r(\mathbf{P})$ is semilinear, then there exists an ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model that has constant running time and constant query complexity.*

Proof. Let $r := r_9(\epsilon)$ be the locality radius of $\mathbf{P}$ for ϵ; let $n_{\min} := n_{\min}(\epsilon)$, $n_{\max} := n_{\max}(\epsilon)$, $f := f(\epsilon)$, and $\mu := \mu(\epsilon)$ be as in Theorem 12; and let $s = \mathrm{c}(r)^2/\mu^2 \cdot \ln(20\,\mathrm{c}(r))$. Assume that the set $\mathrm{h}_r(\mathbf{P})$ is semilinear. Given oracle access to a σ-db $\mathcal{D} \in \mathbf{C}$ and $|D| = n$ as an input, the ϵ-tester proceeds as follows:

(1) If $n \leq n_{\max}$, do a full check of $\mathcal{D}$ and decide if $\mathcal{D} \in \mathbf{P}$.
(2) Run $\mathrm{EstimateFrequencies}_{r,s}$ and let $\bar{v}$ be the resulting vector.
(3) If there exists a $\mathcal{D}' \in \mathbf{P}$ where $n_{\min} \leq |D'| \leq n_{\max}$ and $\|\bar{v} - \operatorname{dv}_r(\mathcal{D}')\|_1 \leq f$, then accept; otherwise reject.

The running time and query complexity of the above tester are constant as $n_{\max}$ is a constant (it only depends on $\mathbf{P}$, d, and ϵ) and $\mathrm{EstimateFrequencies}_{r,s}$ runs in constant time and makes a constant number of queries.

For correctness, first assume $\mathcal{D} \in \mathbf{P}$. By Theorem 12 there exists a σ-db $\mathcal{D}' \in \mathbf{P}$ such that $n_{\min} \leq |D'| \leq n_{\max}$ and $\| \operatorname{dv}_r(\mathcal{D}) - \operatorname{dv}_r(\mathcal{D}')\|_1 \leq f - \mu$. By Lemma 18 with probability at least $9/10$, $\|\bar{v} - \operatorname{dv}_r(\mathcal{D})\|_1 \leq \mu$ and therefore $\|\bar{v} - \operatorname{dv}_r(\mathcal{D}')\|_1 \leq f$. Hence, with probability at least $9/10$ the tester will accept.

Now assume $\mathcal{D}$ is ϵ-far from $\mathbf{P}$. By Theorem 12 for every $\mathcal{D}' \in \mathbf{P}$ with $n_{\min} \leq |D'| \leq n_{\max}$, we have $\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 > f + \mu$. By Lemma 18 with probability at least 9/10, $\|\bar{v} - \mathrm{dv}_r(\mathcal{D})\|_1 \leq \mu$, and therefore for every $\mathcal{D}' \in \mathbf{P}$ with $n_{\min} \leq |D'| \leq n_{\max}$, $\|\bar{v} - \mathrm{dv}_r(\mathcal{D}')\|_1 > f$. Hence, with probability at least 9/10 the tester will reject. □

Combining Theorems 11 and 19, we obtain the following as a corollary.

Theorem 20. *Let* $\mathbf{C}$ *be closed under removing tuples and let* $\mathbf{P} \subseteq \mathbf{C}$ *be a property that is hyperfinite on* $\mathbf{C}$. *If for each* $r \in \mathbb{N}$ *the set* $\mathrm{h}_r(\mathbf{P})$ *is semilinear, then* $\mathbf{P}$ *is uniformly testable on* $\mathbf{C}$ *in constant time in the* $\mathrm{BDRD}_{+/-}$ *model.*

Combining the above theorem (Theorem 20), Lemma 1, and the fact that $\mathbf{C}_d^t$ is hyperfinite [4, 19] (and so any property is hyperfinite on $\mathbf{C}_d^t$), we obtain the following as a corollary.

Theorem 21. *Every property* $\mathbf{P}$ *definable by a CMSO sentence on* $\mathbf{C}_d^t$ *is uniformly testable on* $\mathbf{C}_d^t$ *with constant time complexity in the* $\mathrm{BDRD}_{+/-}$ *model.*

6 EVERY HYPERFINITE PROPERTY THAT IS CLOSE TO HAVING SEMILINEAR NEIGHBORHOOD HISTOGRAMS IS CONSTANT TIME TESTABLE

We begin this section by defining the notion of δ-indistinguishability, which is based on the definition of indistinguishability in the dense graph model given in [2]. We then prove that any hyperfinite property that, for every $\delta \in (0, 1]$, is δ-indistinguishable from a property whose r-histograms are semilinear is constant time testable in the $\mathrm{BDRD}_{+/-}$ model (Theorem 25).

Definition 22 (δ-indistinguishable). Let $\delta \in (0, 1]$. Two properties $\mathbf{P}$ and $\mathbf{Q}$ are called *δ-indistinguishable* if there exists $N := N_{22}(\delta) \in \mathbb{N}$ that satisfies the following. For every σ-db $\mathcal{D} \in \mathbf{P}$ with $|D| = n \geq N$ elements there exists a σ-db $\mathcal{D}' \in \mathbf{Q}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{n, |D'|\}$; and for every σ-db $\mathcal{D} \in \mathbf{Q}$ with $|D| = n \geq N$ elements there exists a σ-db $\mathcal{D}' \in \mathbf{P}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{n, |D'|\}$.

The following two lemmas will be useful in the proof of Theorem 25.

Lemma 23. *Let* $\mathcal{D}$ *and* $\mathcal{D}'$ *be two dbs and let* $\delta \in (0, 1]$. *If* $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{|D|, |D'|\}$, *then for any* $r \in \mathbb{N}$, $\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 \leq 3\delta c d^{r+2}$ *and* $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq 2\delta d^{r+2} \min\{|D|, |D'|\}$, *where* $c := c(r)$.

Proof. Let $\delta \in (0, 1]$ and let $r \in \mathbb{N}$. Let us assume that $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{|D|, |D'|\}$. The distance between $\mathcal{D}$ and $\mathcal{D}'$ is the minimum number of modifications needed to make $\mathcal{D}$ and $\mathcal{D}'$ isomorphic. The four different types of modifications allowed are (1) inserting a new element, (2) deleting an element, (3) inserting a new tuple, and (4) deleting a tuple. If a new element is added to $\mathcal{D}$ or $\mathcal{D}'$, then no existing elements' r-type is changed. If an element is deleted from $\mathcal{D}$ or $\mathcal{D}'$, then the r-type of any element at distance at most r from the deleted element could have changed. If a tuple $\bar{a}$ is inserted or deleted from $\mathcal{D}$ or $\mathcal{D}'$, then the r-type of any element that is at distance at most r from every element in $\bar{a}$ could have changed. Hence, since the number of elements in the r-neighborhood of an element is at most d^{r+1} (cf., e.g., Lemma 3.2 (a) of [8]), every modification to $\mathcal{D}$ or $\mathcal{D}'$ could change the r-type of at most d^{r+1} many elements. Therefore, $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq 2\delta d^{r+2} \min\{|D|, |D'|\}$, as required.

By definition,

$$\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 = \sum_{i=1}^{c} \left| \frac{\mathrm{h}_r(\mathcal{D})[i]}{|D|} - \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D'|} \right|.$$

Let $i \in [c]$; then, since $\| \mathrm{h}_r(\mathcal{D}) - \mathrm{h}_r(\mathcal{D}')\|_1 \leq 2\delta d^{r+2} \min\{|D|, |D'|\}$,

$$\begin{aligned}\frac{\mathrm{h}_r(\mathcal{D})[i]}{|D|} - \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D'|} &\leq \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D|} + \frac{2\delta d^{r+2}\min\{|D|, |D'|\}}{|D|} - \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D'|} \\ &\leq \mathrm{h}_r(\mathcal{D}')[i]\left(\frac{1}{|D|} - \frac{1}{|D'|}\right) + 2\delta d^{r+2}.\end{aligned}$$

Then, since $|D'| \leq |D|(1+\delta d)$ (as $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{|D|, |D'|\}$), we have $|D| \geq |D'|/(1+\delta d)$ and hence

$$\mathrm{h}_r(\mathcal{D}')[i]\left(\frac{1}{|D|} - \frac{1}{|D'|}\right) + 2\delta d^{r+2} \leq \frac{\mathrm{h}_r(\mathcal{D}')[i]\delta d}{|D'|} + 2\delta d^{r+2}.$$

Similarly,

$$\begin{aligned}\frac{\mathrm{h}_r(\mathcal{D})[i]}{|D|} - \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D'|} &\geq \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D|} - \frac{2\delta d^{r+2}\min\{|D|, |D'|\}}{|D|} - \frac{\mathrm{h}_r(\mathcal{D}')[i]}{|D'|} \\ &\geq \mathrm{h}_r(\mathcal{D}')[i]\left(\frac{1}{|D|} - \frac{1}{|D'|}\right) - 2\delta d^{r+2} \\ &\geq \frac{-\mathrm{h}_r(\mathcal{D}')[i]\delta d}{|D'|} - 2\delta d^{r+2}\end{aligned}$$

since $|D'| \geq |D|(1 - \delta d)$ and hence $|D| \leq |D'|/(1 - \delta d)$. Hence,

$$\begin{aligned}\| \mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\|_1 &\leq \sum_{i=1}^{c}\left(\frac{\mathrm{h}_r(\mathcal{D}')[i]\delta d}{|D'|} + 2\delta d^{r+2}\right) \\ &= \delta d + 2\delta c d^{r+2} \\ &\leq 3\delta c d^{r+2},\end{aligned}$$

as required. □

Lemma 24. *Let $D_1, D_2, D_3 \in \mathbb{N}$ and let $x \in \mathbb{R}$ such that $x \geq 0$. If $|D_1 - D_2| \leq x \min\{D_1, D_2\}$, then*

$$(1 - x)\min\{D_1, D_3\} \leq \min\{D_2, D_3\} \leq (1 + x)\min\{D_1, D_3\}.$$

Proof. Let us assume that $|D_1 - D_2| \leq x \min\{D_1, D_2\}$. To prove that $\min\{D_2, D_3\} \leq (1 + x)\min\{D_1, D_3\}$, we shall consider the cases where out of D_1, D_2, and D_3, (1) D_1 is the smallest, (2) D_2 is the smallest, and (3) D_3 is the smallest. In cases (2) and (3), since $(1 + x) \geq 1$, the inequality holds. For case (1), $\min\{D_2, D_3\} \leq D_2 \leq D_1 + x\min\{D_1, D_2\} = (1 + x)\min\{D_1, D_3\}$ since $|D_1 - D_2| \leq x\min\{D_1, D_2\}$ and D_1 is the smallest.

To prove that $(1 - x)\min\{D_1, D_3\} \leq \min\{D_2, D_3\}$, we shall again consider the above three cases. In cases (1) and (3), since $1 - x \leq 1$, the inequality holds. For case (2), $\min\{D_2, D_3\} = D_2 \geq (1 - x)D_1 \geq (1 - x)\min\{D_1, D_3\}$ since $|D_1 - D_2| \leq x\min\{D_1, D_2\} \leq xD_1$. □

We now prove our main result of this section.

Theorem 25. *Let $\mathbf{C}$ be closed under removing tuples and let $\mathbf{P} \subseteq \mathbf{C}$ be a property that is hyperfinite on $\mathbf{C}$. If for every $\delta \in (0, 1]$ there exists a property $\mathbf{Q}_\delta \subseteq \mathbf{C}$ such that*

(1) *$\mathbf{P}$ and $\mathbf{Q}_\delta$ are δ-indistinguishable, and*
(2) *for every $r \in \mathbb{N}$, $\mathrm{h}_r(\mathbf{Q}_\delta)$ is semilinear,*

then $\mathbf{P}$ is uniformly testable on $\mathbf{C}$ in constant time.

Proof. Let $\epsilon \in (0,1]$. We shall prove that there exists an ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ that runs in constant time and has constant query complexity. By Theorem 11, $\mathbf{P}$ is local on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model. Let $N_{\mathbf{P}} := N_9(\epsilon/4)$, $\lambda_{\mathbf{P}} := \lambda_9(\epsilon/4)$, and $r_{\mathbf{P}} := r_9(\epsilon/4)$ be as in Definition 9 for $\mathbf{P}$ and $\epsilon/4$. Let $c := c(r_{\mathbf{P}})$. Let $\delta = \min\{\epsilon/6d, \lambda_{\mathbf{P}}/40c^2d^{r+2}\}$ and let $\mathbf{Q}_\delta \subseteq \mathbf{C}$ be a property that is δ-indistinguishable from $\mathbf{P}$ and for every $r \in \mathbb{N}$, $\mathrm{h}_r(\mathbf{Q}_\delta)$ is semilinear.

Let $r := r_{\mathbf{P}}$, $\lambda := \lambda_{\mathbf{P}}/2$ and $N := (N_{\mathbf{P}}+1)(N_{22}(\delta)+1)(1+\epsilon d/4)$, where $N_{22}(\delta)$ is as in Definition 22 for $\mathbf{P}$ and $\mathbf{Q}_\delta$.

Claim 26. *The property $\mathbf{Q}_\delta$ is $\epsilon/2$-local on $\mathbf{C}$ in the $\mathrm{BDRD}_{+/-}$ model with $r_9(\epsilon/2) = r$, $\lambda_9(\epsilon/2) = \lambda$, and $N_9(\epsilon/2) = N$.*

Proof. Let $\mathcal{D}_{\mathbf{Q}} \in \mathbf{Q}_\delta$ and let $\mathcal{D}_{\mathbf{C}} \in \mathbf{C}$ such that $|D_{\mathbf{Q}}| \geq N$ and $|D_{\mathbf{C}}| \geq N$. Let us assume that $\|\mathrm{h}_r(\mathcal{D}_{\mathbf{Q}}) - \mathrm{h}_r(\mathcal{D}_{\mathbf{C}})\|_1 \leq \lambda \min\{|D_{\mathbf{Q}}|, |D_{\mathbf{C}}|\}$. We need to prove that $\mathcal{D}_{\mathbf{C}}$ is $\epsilon/2$-close to $\mathbf{Q}_\delta$.

Since $\mathbf{P}$ and $\mathbf{Q}_\delta$ are δ-indistinguishable and $N \geq N_{22}(\delta)$, there exists $\mathcal{D}_{\mathbf{P}} \in \mathbf{P}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{P}}, \mathcal{D}_{\mathbf{Q}}) \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}$. By Lemma 23, $\|\mathrm{h}_r(\mathcal{D}_{\mathbf{P}}) - \mathrm{h}_r(\mathcal{D}_{\mathbf{Q}})\| \leq 2\delta d^{r+2} \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}$. Hence,

$$\|\mathrm{h}_r(\mathcal{D}_{\mathbf{P}}) - \mathrm{h}_r(\mathcal{D}_{\mathbf{C}})\| \leq \lambda \min\{|D_{\mathbf{Q}}|, |D_{\mathbf{C}}|\} + 2\delta d^{r+2} \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}.$$

We have $||D_{\mathbf{P}}| - |D_{\mathbf{Q}}|| \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}$ (as $\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{P}}, \mathcal{D}_{\mathbf{Q}}) \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}$) and we have $||D_{\mathbf{C}}| - |D_{\mathbf{Q}}|| \leq \lambda \min\{|D_{\mathbf{C}}|, |D_{\mathbf{Q}}|\}$ (as $\|\mathrm{h}_r(\mathcal{D}_{\mathbf{Q}}) - \mathrm{h}_r(\mathcal{D}_{\mathbf{C}})\|_1 \leq \lambda \min\{|D_{\mathbf{Q}}|, |D_{\mathbf{C}}|\}$). By Lemma 24, $\min\{|D_{\mathbf{Q}}|, |D_{\mathbf{C}}|\} \leq (1+\delta d)\min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\}$ (where $D_1 = |D_{\mathbf{P}}|$, $D_2 = |D_{\mathbf{Q}}|$, and $D_3 = |D_{\mathbf{C}}|$) and $\min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\} \leq (1+\lambda)\min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\}$ (where $D_1 = |D_{\mathbf{C}}|$, $D_2 = |D_{\mathbf{Q}}|$, and $D_3 = |D_{\mathbf{P}}|$).

Therefore,

$$\begin{aligned}\|\mathrm{h}_r(\mathcal{D}_{\mathbf{P}}) - \mathrm{h}_r(\mathcal{D}_{\mathbf{C}})\| &\leq \lambda(1+\delta d)\min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\} + 2\delta d^{r+2}(1+\lambda)\min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\} \\ &\leq (\lambda + 5\delta d^{r+2})\min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\} \\ &\leq \lambda_{\mathbf{P}} \min\{|D_{\mathbf{P}}|, |D_{\mathbf{C}}|\}\end{aligned}$$

by the choice of λ and δ. Since $\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{P}}, \mathcal{D}_{\mathbf{Q}}) \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\}$, we have $|D_{\mathbf{P}}| \geq |D_{\mathbf{Q}}| - \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathbf{Q}}|\} \geq |D_{\mathbf{Q}}| - \delta d |D_{\mathbf{P}}|$. Hence,

$$|D_{\mathbf{P}}| \geq \frac{|D_{\mathbf{Q}}|}{1+\delta d} \geq \frac{N}{1+\epsilon d/4} \geq N_{\mathbf{P}}$$

by the choice of δ and N. Therefore, since $\mathbf{P}$ is local on $\mathbf{C}$, $r = r_{\mathbf{P}}$, and $|D_{\mathbf{P}}| \geq N_{\mathbf{P}}$, $\mathcal{D}_{\mathbf{C}}$ is $\epsilon/4$-close to $\mathbf{P}$.

Since $\mathcal{D}_{\mathbf{C}}$ is $\epsilon/4$-close to $\mathbf{P}$, there exists $\mathcal{D}'_{\mathbf{P}} \in \mathbf{P}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{C}}, \mathcal{D}'_{\mathbf{P}}) \leq \epsilon d \min\{|D_{\mathbf{C}}|, |D'_{\mathbf{P}}|\}/4$ (which implies $||D_{\mathbf{C}}| - |D'_{\mathbf{P}}|| \leq \epsilon d \min\{|D_{\mathbf{C}}|, |D'_{\mathbf{P}}|\}/4$). Since $|D'_{\mathbf{P}}| \geq |D_{\mathbf{C}}| - \epsilon d \min\{|D_{\mathbf{C}}|, |D'_{\mathbf{P}}|\}/4 \geq |D_{\mathbf{C}}| - \epsilon d |D'_{\mathbf{P}}|/4$, we have

$$|D'_{\mathbf{P}}| \geq \frac{|D_{\mathbf{C}}|}{(1+\epsilon d/4)} \geq \frac{N}{(1+\epsilon d/4)} \geq N_{22}(\delta)$$

by the choice of N. Therefore, there exists $\mathcal{D}'_{\mathbf{Q}} \in \mathbf{Q}_\delta$ such that $\mathrm{dist}_{+/-}(\mathcal{D}'_{\mathbf{P}}, \mathcal{D}'_{\mathbf{Q}}) \leq \delta d \min\{|D'_{\mathbf{P}}|, |D'_{\mathbf{Q}}|\}$ (which implies $||D'_{\mathbf{Q}}| - |D'_{\mathbf{P}}|| \leq \delta d \min\{|D'_{\mathbf{Q}}|, |D'_{\mathbf{P}}|\}$) . Hence,

$$\begin{aligned}\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{C}}, \mathcal{D}'_{\mathbf{Q}}) &\leq \frac{\epsilon d \min\{|D_{\mathbf{C}}|, |D'_{\mathbf{P}}|\}}{4} + \delta d \min\{|D'_{\mathbf{P}}|, |D'_{\mathbf{Q}}|\} \\ &\leq \frac{\epsilon d(1+\delta d)\min\{|D_{\mathbf{C}}|, |D'_{\mathbf{Q}}|\}}{4} + \delta d\left(1 + \frac{\epsilon d}{4}\right)\min\{|D_{\mathbf{C}}|, |D'_{\mathbf{Q}}|\}\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{\epsilon d}{4} + \frac{\epsilon \delta d^2}{2} + \delta d\right) \min\{|D_{\mathrm{C}}|, |D'_{\mathrm{Q}}|\} \\
&\leq \left(\frac{\epsilon d}{4} + \frac{\epsilon^2 d}{12} + \frac{\epsilon}{6}\right) \min\{|D_{\mathrm{C}}|, |D'_{\mathrm{Q}}|\} \\
&\leq \left(\frac{\epsilon d}{4} + \frac{\epsilon d}{12} + \frac{\epsilon d}{6}\right) \min\{|D_{\mathrm{C}}|, |D'_{\mathrm{Q}}|\} \\
&= \frac{\epsilon d}{2} \min\{|D_{\mathrm{C}}|, |D'_{\mathrm{Q}}|\}
\end{aligned}
$$

by Lemma 24 and the choice of δ. Therefore, $\mathcal{D}_{\mathrm{C}}$ is $\epsilon/2$-close to $\mathbf{Q}_\delta$ as required. □

By Claim 26 and Theorem 19 there exists an $\epsilon/2$-tester for $\mathbf{Q}_\delta$ on $\mathbf{C}$ that runs in constant time and has constant query complexity. Let $\mu := \mu(\epsilon/2)$, $f := f(\epsilon/2)$, $n_{\min} := n_{\min}(\epsilon/2)$, and $n_{\max} := n_{\max}(\epsilon/2)$ be as in Theorem 12 for $\mathbf{Q}_\delta$ and $\epsilon/2$. Note that by Claim 26 the locality radius and disc proximity of $\mathbf{Q}_\delta$ for $\epsilon/2$ is $r_{\mathbf{P}}$ and $\lambda_{\mathbf{P}}/2$, respectively; therefore, by Theorem 12, $\mu = \lambda_{\mathbf{P}}/12c$. Let $\pi_{\epsilon/2}$ be the $\epsilon/2$-tester for $\mathbf{Q}_\delta$ on $\mathbf{C}$ from the proof of Theorem 19, but in $\pi_{\epsilon/2}$ let us increase the number of elements sampled in the second step to $s = \mathrm{c}^2/(\mu - 3c\delta d^{r+2})^2 \cdot \ln(20\,\mathrm{c})$. Note that $\mu - 3c\delta d^{r+2} \in (0, 1)$ by the choice of δ. Then, given oracle access to a σ-db $\mathcal{D} \in \mathbf{C}$ and $|D| = n$ as an input, the ϵ-tester for $\mathbf{P}$ on $\mathbf{C}$ proceeds as follows:

(1) If $n < N_{22}(\delta)(1 + \epsilon d/2)$, do a full check of $\mathcal{D}$ and decide if $\mathcal{D} \in \mathbf{P}$.
(2) Run $\pi_{\epsilon/2}$ on $\mathcal{D}$ and accept if $\pi_{\epsilon/2}$ accepts and reject otherwise.

Clearly the above tester runs in constant time and has constant query complexity.

For correctness, first assume that $\mathcal{D} \in \mathbf{P}$ and $|D| = n > n_{\max}$ and $n \geq N_{22}(\delta)(1 + \epsilon d/2)$. As $\mathbf{P}$ and $\mathbf{Q}_\delta$ are δ-indistinguishable, there exists a σ-db $\mathcal{D}' \in \mathbf{Q}_\delta$ such that $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{n, |D'|\}$. By Lemma 23, $\|\,\mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}')\| \leq 3c\delta d^{r+2}$. By Theorem 12 there exists a σ-db $\mathcal{D}_0 \in \mathbf{Q}_\delta$ such that $n_{\min} \leq |D_0| \leq n_{\max}$ and $\|\,\mathrm{dv}_r(\mathcal{D}') - \mathrm{dv}_r(\mathcal{D}_0)\|_1 \leq f - \mu$. Hence, $\|\,\mathrm{dv}_r(\mathcal{D}) - \mathrm{dv}_r(\mathcal{D}_0)\|_1 \leq f - \mu + 3c\delta d^{r+2}$. By Lemma 18 and the choice of s with probability at least 9/10, $\|\bar{v} - \mathrm{dv}_r(\mathcal{D})\|_1 \leq \mu - 3c\delta d^{r+2}$ and therefore $\|\bar{v} - \mathrm{dv}_r(\mathcal{D}_0)\|_1 \leq f$. Hence, with probability at least 9/10 the tester will accept.

Now assume $\mathcal{D}$ is ϵ-far from $\mathbf{P}$ and $|D| \geq N_{22}(\delta)(1 + \epsilon d/2)$. We will show that $\mathcal{D}$ is $\epsilon/2$-far from $\mathbf{Q}_\delta$. Let $\mathcal{D}_{\mathrm{Q}} \in \mathbf{Q}_\delta$. If $||D_{\mathrm{Q}}| - |D|| > \epsilon d \min\{|D_{\mathrm{Q}}|, |D|\}/2$, then $\mathcal{D}$ is $\epsilon/2$-far from $\mathcal{D}_{\mathrm{Q}}$. So let us assume that $||D_{\mathrm{Q}}| - |D|| \leq \epsilon d \min\{|D_{\mathrm{Q}}|, |D|\}/2$. This implies that $|D|/(1 + \epsilon d/2) \leq |D_{\mathrm{Q}}| \leq |D|(1 + \epsilon d/2)$. Hence, since $|D| \geq N_{22}(\delta)(1 + \epsilon d/2)$, $|D_{\mathrm{Q}}| \geq N_{22}(\delta)$. Therefore, there exists a σ-db $\mathcal{D}_{\mathbf{P}} \in \mathbf{P}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}_{\mathbf{P}}, \mathcal{D}_{\mathrm{Q}}) \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\}$ (and therefore $||D_{\mathrm{Q}}| - |D_{\mathbf{P}}|| \leq \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\}$). Since $\mathcal{D}$ is ϵ-far from $\mathbf{P}$, $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}_{\mathbf{P}}) > \epsilon d \min\{|D|, |D_{\mathbf{P}}|\}$ and so $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}_{\mathrm{Q}}) > \epsilon d \min\{|D|, |D_{\mathbf{P}}|\} - \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\}$. By Lemma 24 (with $D_1 = |D_{\mathrm{Q}}|$, $D_2 = |D_{\mathbf{P}}|$, and $D_3 = |D|$), $\min\{|D|, |D_{\mathbf{P}}|\} \geq (1 - \delta d) \min\{|D|, |D_{\mathrm{Q}}|\}$. Furthermore, we can show that $\min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\} \leq (1 + \epsilon d/2) \min\{|D|, |D_{\mathrm{Q}}|\}$. To see this consider the case when $|D|$ is the smallest out of $|D|$, $|D_{\mathrm{Q}}|$, and $|D_{\mathbf{P}}|$; then

$$\min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\} \leq |D_{\mathrm{Q}}| \leq |D|(1 + \epsilon d/2) = (1 + \epsilon d/2) \min\{|D|, |D_{\mathrm{Q}}|\}.$$

If $|D_{\mathrm{Q}}|$ or $|D_{\mathbf{P}}|$ are the smallest, then the inequality clearly holds. Therefore,

$$
\begin{aligned}
\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}_{\mathrm{Q}}) &> \epsilon d \min\{|D|, |D_{\mathbf{P}}|\} - \delta d \min\{|D_{\mathbf{P}}|, |D_{\mathrm{Q}}|\} \\
&\geq \left(\epsilon d(1 - \delta d) - \delta d \left(1 + \frac{\epsilon d}{2}\right)\right) \min\{|D|, |D_{\mathrm{Q}}|\}
\end{aligned}
$$

$$\geq \left(\epsilon d - \frac{\epsilon}{6} - \frac{\epsilon^2 d}{4}\right) \min\{|D|, |D_{\mathbf{Q}}|\}$$
$$\geq \left(\epsilon d - \frac{\epsilon d}{6} - \frac{\epsilon d}{4}\right) \min\{|D|, |D_{\mathbf{Q}}|\}$$
$$\geq \frac{\epsilon d}{2} \min\{|D|, |D_{\mathbf{Q}}|\}$$

by the choice of δ. Hence, $\mathcal{D}$ is $\epsilon/2$-far from every $\mathcal{D}_{\mathbf{Q}} \in \mathbf{Q}_\delta$ and so $\mathcal{D}$ is $\epsilon/2$-far from $\mathbf{Q}_\delta$. As $\pi_{\epsilon/2}$ is an $\epsilon/2$-tester for $\mathbf{Q}_\delta$ on $\mathbf{C}$, with probability at least 2/3 the tester will reject. □

7 CONSTANT TIME TESTABILITY OF HYPERFINITE HEREDITARY PROPERTIES

A db (or graph) property is called *hereditary* if it is closed under removing elements (or vertices). To the best of our knowledge, in the bounded degree graph model, it has not been shown explicitly that hyperfinite hereditary properties are uniformly (in n) testable in constant time. Benjamini et al. [7] prove that every monotone hyperfinite property is uniformly testable in constant time (in the bounded degree graph model). Their tester starts by testing for hyperfiniteness (which they show can be done in constant time). Newman and Sohler [25] prove that every hyperfinite property is non-uniformly (in n) testable. We are interested in obtaining testers that are uniform in n *and* run in constant time. Furthermore, Hassidim et al. [19] show that it is possible to approximate the distance to non-degenerate hereditary properties for hyperfinite graphs in constant time.

With methods similar to [7] and [12] it can be shown that every hereditary hyperfinite property is uniformly testable in constant time in the BDRD model and we will sketch this below.

Theorem 27. *Every hyperfinite hereditary property is uniformly testable on* $\mathbf{C}$ *in constant time in the* BDRD *model.*

By Lemma 4 and Theorem 27 we immediately get that every hyperfinite hereditary property is uniformly testable in constant time in the $\mathrm{BDRD}_{+/-}$ model.

Theorem 28. *Every hyperfinite hereditary property* $\mathbf{P} \subseteq \mathbf{C}$ *is uniformly testable on* $\mathbf{C}$ *in constant time in the* $\mathrm{BDRD}_{+/-}$ *model.*

In this section we will sketch a proof of Theorem 27 that follows closely to the proof given in [7]. We will then give an alternative proof of Theorem 28. In our alternative proof we show that every hereditary hyperfinite property is close to having semilinear neighborhood histograms and hence by Theorem 25 is uniformly testable in constant time in the $\mathrm{BDRD}_{+/-}$ model.

7.1 A Proof of Theorem 27

First we start with some definitions that are based on those used in [7].

Let $b \in \mathbb{N}$ and let $\Psi(b)$ be the set of all non-isomorphic connected σ-dbs of size at most b. For each $S \subseteq \Psi(b)$, let $\mathcal{D}(S)$ be the disjoint union of the σ-dbs in S. Let $\mathbf{P}$ be some hereditary property. Then let $\Phi_{\mathbf{P}}(S)$ be the smallest integer g such that the σ-db obtained by taking g disjoint copies of $\mathcal{D}(S)$ is not in $\mathbf{P}$. If no such integer exists (i.e., every σ-db that only contains connected components isomorphic to those in S is in $\mathbf{P}$), then $\Phi_{\mathbf{P}}(S) = \infty$.

Definition 29. For a fixed hereditary property $\mathbf{P}$ and $b \in \mathbb{N}$, let $\Pi_{\mathbf{P}}^b = \{S \subseteq \Psi(b) \mid \Phi_{\mathbf{P}}(S) < \infty\}$. We then define the function $\Phi_{\mathbf{P}} : \mathbb{N} \mapsto \mathbb{N}$ as follows:

$$\Phi_{\mathbf{P}}(b) = \begin{cases} 0 & \text{if } \Pi_{\mathbf{P}}^b = \emptyset \\ \max\limits_{S \subseteq \Pi_{\mathbf{P}}^b} \Phi_{\mathbf{P}}(S) & \text{otherwise.} \end{cases}$$

Note that the function $\Phi_{\mathbf{P}}(b)$ is well defined as the set $\Psi(b)$ is finite. We will illustrate Definition 29 in the following example.

Example 30. Let $\mathbf{P}$ be the property containing all bounded degree chordal graphs (a graph is *chordal* if every cycle of length four or greater has a chord, where a *chord* of a cycle is an edge that is not in the edge set of the cycle but has endpoints in the cycle). A graph $\mathcal{G}$ is chordal if and only if $\mathcal{G}$ does not contain a cycle of length four or greater as an induced subgraph. Therefore, the set of minimal forbidden induced subgraphs of $\mathbf{P}$ is the set of cycles of length four or greater; i.e., every forbidden induced subgraph has one connected component. For any $b \in \mathbb{N}$ and $S \subseteq \Psi(b)$, if one of the graphs in S contains a cycle of length four or greater as an induced subgraph, then $\Phi_{\mathbf{P}}(S) = 1$; otherwise $\Phi_{\mathbf{P}}(S) = \infty$. Hence, if $b > 3$, then $\Phi_{\mathbf{P}}(b) = 1$; otherwise $\Phi_{\mathbf{P}}(b) = 0$.

Sketch of the proof of Theorem 27. Let $\epsilon \in (0, 1]$ and let $\mathbf{P}$ be a hyperfinite hereditary property on $\mathbf{C}$. Let $\epsilon_0 = \epsilon_0(\epsilon)$ be a carefully chosen constant and let k be such that any db in $\mathbf{P}$ is (ϵ_0, k)-hyperfinite. Let us start by describing the ϵ-tester. Let $\mathcal{D}$ be the input db. The tester starts by deciding correctly with high probability whether $\mathcal{D}$ is (ϵ_0, k)-hyperfinite or not $(\epsilon/2, k)$-hyperfinite (ϵ_0 is chosen in such a way that $\mathcal{D}$ cannot be both (ϵ_0, k)-hyperfinite and not $(\epsilon/2, k)$-hyperfinite by the choice of ϵ_0). This can be done in constant time and with constant query complexity in the BDRD model (extending methods in [7]). This is an ϵ-tester for the property of being (ϵ_0, k)-hyperfinite since if $\mathcal{D}$ is ϵ-far from being (ϵ_0, k)-hyperfinite, it is not $(\epsilon/2, k)$-hyperfinite. If $\mathcal{D}$ is declared to be not (ϵ_0, k)-hyperfinite, then the tester rejects. If $\mathcal{D}$ is declared to be $(\epsilon/2, k)$-hyperfinite, then the tester samples a constant number $m = m(\epsilon)$ of elements from $\mathcal{D}$ and for each element the tester explores its k-neighborhood. If the induced sub-database of $\mathcal{D}$ on the union of the k-neighborhoods of the sampled elements is not in $\mathbf{P}$, then the tester rejects. Otherwise, the tester accepts. This can be done with constant running time and constant query complexity.

Now to prove correctness (which follows closely to that in [7]), let us assume that $\mathcal{D} \in \mathbf{P}$. By the choice of k, $\mathcal{D}$ is (ϵ_0, k)-hyperfinite and so with high probability will be accepted in the first step of the tester. Then, since $\mathbf{P}$ is hereditary, the second step of the tester will accept with probability 1.

Let us now assume that $\mathcal{D}$ is ϵ-far from $\mathbf{P}$. Let us assume that $\mathcal{D}$ is $(\epsilon/2, k)$-hyperfinite (otherwise the tester would reject with high probability). Let $\mathcal{D}'$ be the db that is formed from $\mathcal{D}$ by removing the minimum number of tuples required (at most $\epsilon n/2$ tuples) such that every connected component in $\mathcal{D}'$ is of size at most k. Note that $\mathcal{D}'$ is at least $\epsilon/2$-far from $\mathbf{P}$. Let $S \subseteq \Psi(k)$ be such that each $\mathcal{A} \in \Psi(k)$ is in S if and only if there are at least $\epsilon n/4k|\Psi(k)|$ connected components in $\mathcal{D}'$ isomorphic to $\mathcal{A}$. We then let $\mathcal{D}''$ be the db formed from $\mathcal{D}'$ as follows. For every $\mathcal{A} \in \Psi(k)$, if $\mathcal{A} \notin S$, then remove every tuple in $\mathcal{D}'$ that is in a connected component isomorphic to $\mathcal{A}$. It is easy to see that the total number of tuples removed is at most $\epsilon dn/4$ and therefore $\mathcal{D}''$ is $\epsilon/4$-far from $\mathbf{P}$. Since $\mathcal{D}'' \notin \mathbf{P}$ and $\mathbf{P}$ is hereditary, $\Phi_{\mathbf{P}}(S) \leq \Phi_{\mathbf{P}}(k) < \infty$. Since each $\mathcal{A} \in S$ appears at least $\epsilon n/4k|\Psi(k)|$ times in $\mathcal{D}''$ and $D'' = D$, we can choose m, the number of elements the tester samples, carefully to ensure that with high probability for each $\mathcal{A} \in S$ the tester samples at least $\Phi_{\mathbf{P}}(S)$ elements from connected components in $\mathcal{D}''$ that are isomorphic to $\mathcal{A}$. Furthermore, if we assume n is greater than some function of ϵ, then with high probability each sampled element is from a distinct connected component (in $\mathcal{D}''$) and their $k + 1$-neighborhoods don't intersect (in $\mathcal{D}$). Let $a_1, \dots a_m$ be the elements sampled in the tester. Let $\mathcal{D}_0$ be the induced sub-database of $\mathcal{D}$ on the union of the k-neighborhoods of $a_1, \dots a_m$ and let $\mathcal{D}_0''$ be the union of the connected components of $\mathcal{D}''$ that contain an element a_i. With high probability, by definition, $\mathcal{D}_0'' \notin \mathbf{P}$. Let $a \in D$. It is easy to see that the connected component in $\mathcal{D}''$ containing a is an induced sub-database of the k-neighborhood of a in $\mathcal{D}$ (since $\mathcal{D}'$ was formed with the minimum required number of tuple deletions). Therefore, if none of the $k + 1$-neighborhoods of $a_1, \dots a_m$ in $\mathcal{D}$ intersect (which happens with high probability), $\mathcal{D}_0''$ is an induced sub-database of $\mathcal{D}_0$. Finally, since $\mathbf{P}$ is hereditary and with high probability $\mathcal{D}_0'' \notin \mathbf{P}$, with high probability $\mathcal{D}_0 \notin \mathbf{P}$ and the tester rejects.

7.2 An Alternative Proof of Theorem 28

We will show that for every hyperfinite hereditary property $\mathbf{P}$ and $\delta \in (0,1]$ there exists a property $\mathbf{Q}$ that has a semilinear set of r-histograms and is δ-indistinguishable from $\mathbf{P}$.

Lemma 31. *Let $\mathbf{P} \subseteq \mathbf{C}$ be a hyperfinite hereditary property and let $\delta \in (0,1]$. Let ρ be the function such that $\mathbf{P}$ is ρ-hyperfinite on $\mathbf{C}$ and let*

$$b := \rho\left(\frac{\delta d}{2(1+\delta d)}\right).$$

Let $\mathbf{Q} \subseteq \mathbf{P}$ be the property such that for every $\mathcal{D} \in \mathbf{P}$, $\mathcal{D} \in \mathbf{Q}$ if and only if all connected components in $\mathcal{D}$ are of size at most b and for each $\mathcal{A} \in \Psi(b)$, $\mathcal{D}$ has either 0 or at least $\Phi_{\mathbf{P}}(b)$ connected components isomorphic to $\mathcal{A}$. Then

(1) *$\mathbf{P}$ and $\mathbf{Q}$ are δ-indistinguishable, and*
(2) *for every $r \in \mathbb{N}$, $\mathrm{h}_r(\mathbf{Q})$ is semilinear.*

Proof. Let us start by proving 1 of the lemma statement. Let

$$N := \frac{2(1+\delta d) \cdot \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b}{\delta d}.$$

We will show that $\mathbf{P}$ and $\mathbf{Q}$ are δ-indistinguishable with $N_{22}(\delta) = N$. If $\mathcal{D} \in \mathbf{Q}$, then $\mathcal{D} \in \mathbf{P}$ as $\mathbf{Q} \subseteq \mathbf{P}$. Hence, to prove $\mathbf{P}$ and $\mathbf{Q}$ are δ-indistinguishable, we only need to show that for every $\mathcal{D} \in \mathbf{P}$ with $|D| = n \geq N$ there exists a σ-db $\mathcal{D}' \in \mathbf{Q}$ such that $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{n, |D'|\}$. Let $\mathcal{D} \in \mathbf{P}$ with $|D| = n > N$. As $\mathbf{P}$ is ρ-hyperfinite on $\mathbf{C}$, by removing at most $\frac{\delta d n}{2(1+\delta d)}$ tuples from $\mathcal{D}$, we can obtain a σ-db $\mathcal{D}_1 \in \mathbf{C}$ that has connected components of size at most b (by the choice of b). For every tuple $\bar{a}$ that was removed from $\mathcal{D}$ to form $\mathcal{D}_1$, pick one element from $\bar{a}$ and remove it (and as a result any tuple containing that element) from $\mathcal{D}$. Let $\mathcal{D}_2$ be the resulting σ-db. Note that as $\mathbf{P}$ is hereditary, $\mathcal{D}_2 \in \mathbf{P}$. Furthermore, $\mathcal{D}_2$ has connected components of size at most b (as $\mathcal{D}_2$ is a sub-database of $\mathcal{D}_1$) and as at most $\frac{\delta d n}{2(1+\delta d)}$ elements were removed from $\mathcal{D}$ to form $\mathcal{D}_2$,

$$\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}_2) \leq \frac{\delta d n}{2(1+\delta d)} \text{ and } |D_2| \geq \left(1 - \frac{\delta d}{2(1+\delta d)}\right) n = \frac{n(2+\delta d)}{2(1+\delta d)}.$$

Now for every $\mathcal{A} \in \Psi(b)$, if $\mathcal{D}_2$ contains less than $\Phi_{\mathbf{P}}(b)$ many connected components isomorphic to $\mathcal{A}$, remove all such connected components from $\mathcal{D}_2$. Let $\mathcal{D}'$ be the resulting σ-db. As $\mathbf{P}$ is hereditary, $\mathcal{D}' \in \mathbf{P}$, and by the construction of $\mathcal{D}'$, all connected components in $\mathcal{D}'$ are of size at most b and for each $\mathcal{A} \in \Psi(b)$, $\mathcal{D}'$ has either 0 or at least $\Phi_{\mathbf{P}}(b)$ connected components isomorphic to $\mathcal{A}$. Therefore, $\mathcal{D}' \in \mathbf{Q}$. At most $\Phi_{\mathbf{P}}(b) \cdot |\Psi(b)|$ many connected components were removed from $\mathcal{D}_2$ to form $\mathcal{D}'$ and hence,

$$|D'| \geq |D_2| - \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b \geq \frac{n(2+\delta d)}{2(1+\delta d)} - \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b$$

and $\mathrm{dist}_{+/-}(\mathcal{D}_2, \mathcal{D}') \leq \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b$. Therefore,

$$\begin{aligned}
\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') &\leq \frac{\delta d n}{2(1+\delta d)} + \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b \\
&\leq \frac{\delta d n}{2(1+\delta d)} + \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b + \frac{\delta d n}{2} - (1+\delta d) \cdot \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b \\
&= \frac{\delta d n(2+\delta d)}{2(1+\delta d)} - \delta d \cdot \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b \\
&\leq \delta d |D'|,
\end{aligned}$$

as $n \geq N$ and so $\frac{\delta dn}{2} - (1 + \delta d) \cdot \Phi_{\mathbf{P}}(b) \cdot |\Psi(b)| \cdot b \geq 0$. When constructing $\mathcal{D}'$ from $\mathcal{D}$, we only removed elements and hence $|D'| \leq |D|$. Therefore, $\mathrm{dist}_{+/-}(\mathcal{D}, \mathcal{D}') \leq \delta d \min\{|D|, |D'|\}$. This completes the proof that $\mathbf{P}$ and $\mathbf{Q}$ are δ-indistinguishable.

We will now prove 2 of the lemma statement. Let $r \in \mathbb{N}$ and for every $S \subseteq \Psi(b)$ let $\mathbf{Q}_S \subseteq \mathbf{Q}$ be the set of σ-dbs such that for every $\mathcal{D} \in \mathbf{Q}$, $\mathcal{D} \in \mathbf{Q}_S$ if and only if $\mathcal{D}$ contains at least $\Phi_{\mathbf{P}}(b)$ many connected components isomorphic to every $\mathcal{A} \in S$ but does not contain a connected component isomorphic to a σ-db in $\Psi(b) \setminus S$. Note that

$$\mathbf{Q} = \bigcup_{S \subseteq \Psi(b)} \mathbf{Q}_S.$$

We will prove that for every $S \subseteq \Psi(b)$, if $\Phi_{\mathbf{P}}(S) < \infty$, then $\mathbf{Q}_S$ is empty, and if $\Phi_{\mathbf{P}}(S) = \infty$, then $\mathrm{h}_r(\mathbf{Q}_S)$ is a linear set. Since $\mathbf{Q}$ is the union of the sets $\mathbf{Q}_S$, this will imply that $\mathrm{h}_r(\mathbf{Q})$ is a semilinear set.

First, let us prove that for every $S \subseteq \Psi(b)$, if $\Phi_{\mathbf{P}}(S) < \infty$, then $\mathbf{Q}_S$ is empty. For a contradiction assume that for some $S \subseteq \Psi(b)$, $\Phi_{\mathbf{P}}(S) < \infty$ and there exists a σ-db $\mathcal{D} \in \mathbf{Q}_S$. Let $\mathcal{D}'$ be the σ-db that for each $\mathcal{A} \in S$ contains exactly $\Phi_{\mathbf{P}}(b)$ connected components isomorphic to $\mathcal{A}$ and contains no other connected components. By the definition of $\Phi_{\mathbf{P}}(b)$ and as $\Phi_{\mathbf{P}}(S) < \infty$, $\mathcal{D}' \notin \mathbf{P}$. However, $\mathcal{D}'$ is an induced sub-database of $\mathcal{D}$, and since $\mathbf{P}$ is hereditary and $\mathcal{D} \in \mathbf{P}$ (as $\mathbf{Q}_S \subseteq \mathbf{P}$), this implies $\mathcal{D}' \in \mathbf{P}$, which is a contradiction. Hence, for every $S \subseteq \Psi(b)$, if $\Phi_{\mathbf{P}}(S) < \infty$, then $\mathbf{Q}_S$ is empty.

We will now prove that for every $S \subseteq \Psi(b)$, if $\Phi_{\mathbf{P}}(S) = \infty$, then $\mathrm{h}_r(\mathbf{Q}_S)$ is a linear set. Let $S = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_\ell\} \subseteq \Psi(b)$, let $\bar{v} = \sum_{1 \leq i \leq \ell} \Phi_{\mathbf{P}}(b)\, \mathrm{h}_r(\mathcal{D}_i)$, and let

$$M = \{\bar{v} + a_1\, \mathrm{h}_r(\mathcal{D}_1) + a_2\, \mathrm{h}_r(\mathcal{D}_2) + \cdots + a_\ell\, \mathrm{h}_r(\mathcal{D}_\ell) \mid a_1, \ldots, a_\ell \in \mathbb{N}\}.$$

Clearly M is a linear set and we claim that $M = \mathrm{h}_r(\mathbf{Q}_S)$. Let $\bar{u} = \bar{v} + u_1\, \mathrm{h}_r(\mathcal{D}_1) + u_2\, \mathrm{h}_r(\mathcal{D}_2) + \cdots + u_\ell\, \mathrm{h}_r(\mathcal{D}_\ell) \in M$ for some $u_1, \ldots, u_\ell \in \mathbb{N}$ and let $\mathcal{D}$ be the σ-db with exactly $\Phi_{\mathbf{P}}(b) + u_1$ connected components isomorphic to $\mathcal{D}_1$, $\Phi_{\mathbf{P}}(b) + u_2$ connected components isomorphic to $\mathcal{D}_2, \ldots, \Phi_{\mathbf{P}}(b) + u_\ell$ connected components isomorphic to $\mathcal{D}_\ell$, and no other connected components. Clearly $\bar{u} = \mathrm{h}_r(\mathcal{D})$. Then as P is hereditary and $\Phi_{\mathbf{P}}(S) = \infty$, $\mathcal{D} \in \mathbf{P}$, and so by the definition of $\mathbf{Q}_S$, $\mathcal{D} \in \mathbf{Q}_S$. On the other hand, let $\mathcal{D} \in \mathbf{Q}_S$; then by definition for some $u_1, \ldots, u_\ell \in \mathbb{N}$, $\mathcal{D}$ contains exactly $\Phi_{\mathbf{P}}(b) + u_1$ connected components isomorphic to $\mathcal{D}_1$, $\Phi_{\mathbf{P}}(b) + u_2$ connected components isomorphic to $\mathcal{D}_2, \ldots, \Phi_{\mathbf{P}}(b) + u_\ell$ connected components isomorphic to $\mathcal{D}_\ell$, and no other connected components. The r-histogram vector of $\mathcal{D}$ is then $\bar{v} + u_1\, \mathrm{h}_r(\mathcal{D}_1) + u_2\, \mathrm{h}_r(\mathcal{D}_2) + \cdots + u_\ell\, \mathrm{h}_r(\mathcal{D}_\ell)$ and hence $\mathrm{h}_r(\mathcal{D}) \in M$. Therefore, $M = \mathrm{h}_r(\mathbf{Q}_S)$.

We have proven that for every $S \subseteq \Psi(b)$, $\mathrm{h}_r(\mathbf{Q}_S)$ is either empty or a linear set and hence $\mathrm{h}_r(\mathbf{Q})$ is semilinear. □

Combining Theorem 25 and Lemma 31, we obtain Theorem 28 as a corollary.

REFERENCES

[1] Isolde Adler and Frederik Harwath. 2018. Property testing for bounded degree databases. In *35th Symposium on Theoretical Aspects of Computer Science (STACS'18)*, Vol. 96. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 6.

[2] Noga Alon, Eldar Fischer, Michael Krivelevich, and Mario Szegedy. 2000. Efficient testing of large graphs. *Combinatorica* 20, 4 (2000), 451–476.

[3] Noga Alon, Tali Kaufman, Michael Krivelevich, and Dana Ron. 2008. Testing triangle-freeness in general graphs. *SIAM Journal on Discrete Mathematics* 22, 2 (2008), 786–819.

[4] Noga Alon, Paul D. Seymour, and Robin Thomas. 1990. A separator theorem for graphs with an excluded minor and its applications. In *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, Harriet Ortiz (Ed.). ACM, 293–299. https://doi.org/10.1145/100216.100254

[5] Jasine Babu, Areej Khoury, and Ilan Newman. 2016. Every property of outerplanar graphs is testable. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM'16)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[6] Sagi Ben-Moshe, Yaron Kanza, Eldar Fischer, Arie Matsliah, Mani Fischer, and Carl Staelin. 2011. Detecting and exploiting near-sortedness for efficient relational query evaluation. In *Proceedings of the 14th International Conference on Database Theory*. ACM, 256–267.
[7] Itai Benjamini, Oded Schramm, and Asaf Shapira. 2010. Every minor-closed property of sparse graphs is testable. *Advances in Mathematics* 223, 6 (2010), 2200–2218.
[8] Christoph Berkholz, Jens Keppeler, and Nicole Schweikardt. 2018. Answering FO+ MOD queries under updates on bounded degree databases. *ACM Transactions on Database Systems (TODS)* 43, 2 (2018), 7.
[9] Hubie Chen and Yuichi Yoshida. 2019. Testability of homomorphism inadmissibility: Property testing meets database theory. In *Proceedings of the 38th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM, 365–382.
[10] Bruno Courcelle. 1990. Graph rewriting: An algebraic and logic approach. In *Formal Models and Semantics*. Elsevier, 193–242.
[11] Bruno Courcelle and Joost Engelfriet. 2012. *Graph Structure and Monadic Second-order Logic: A Language-theoretic Approach*. Vol. 138. Cambridge University Press.
[12] Artur Czumaj, Asaf Shapira, and Christian Sohler. 2009. Testing hereditary properties of nonexpanding bounded-degree graphs. *SIAM J. Comput.* 38, 6 (2009), 2499–2510.
[13] Arnaud Durand and Etienne Grandjean. 2007. First-order queries on structures of bounded degree are computable with constant delay. *ACM Transactions on Computational Logic (TOCL)* 8, 4 (2007), 21.
[14] Hendrik Fichtenberger, Pan Peng, and Christian Sohler. 2015. On constant-size graphs that preserve the local structure of high-girth graphs. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM'15) (LIPIcs, Vol. 40)*, Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim (Eds.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 786–799. https://doi.org/10.4230/LIPIcs.APPROX-RANDOM.2015.786
[15] Hendrik Fichtenberger, Pan Peng, and Christian Sohler. 2019. Every testable (infinite) property of bounded-degree graphs contains an infinite hyperfinite subproperty. In *Proceedings of the 30th Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 714–726.
[16] Eldar Fischer, Johann A. Makowsky. 2004. On spectra of sentences of monadic second order logic with counting. *Journal of Symbolic Logic* 69, 3 (2004), 617–640.
[17] Jörg Flum and Martin Grohe. 2006. *Parameterized Complexity Theory (Texts in Theoretical Computer Science, an EATCS Series)*. Springer-Verlag, Berlin.
[18] Oded Goldreich and Dana Ron. 2002. Property testing in bounded degree graphs. *Algorithmica* 32, 2 (2002), 302–343.
[19] Avinatan Hassidim, Jonathan A. Kelner, Huy N. Nguyen, and Krzysztof Onak. 2009. Local graph partitions for approximation and testing. In *50th Annual IEEE Symposium on Foundations of Computer Science (FOCS'09)*. IEEE Computer Society, 22–31. https://doi.org/10.1109/FOCS.2009.77
[20] Piotr Indyk, Andrew McGregor, Ilan Newman, and Krzysztof Onak. 2011. Open problems in data streams, property testing, and related topics. In *Bernitoro Workshop on Sublinear Algorithms*.
[21] Hiro Ito, Areej Khoury, and Ilan Newman. 2020. On the characterization of 1-sided error strongly testable graph properties for bounded-degree graphs. *Computational Complexity* 29, 1 (2020), 1.
[22] Wojciech Kazana and Luc Segoufin. 2011. First-order query evaluation on structures of bounded degree. *Logical Methods in Computer Science* 7, 2 (2011), 1–8. https://doi.org/10.2168/LMCS-7(2:20)2011
[23] Leonid Libkin. 2004. *Elements of Finite Model Theory*. Springer.
[24] László Lovász. 2012. *Large Networks and Graph Limits*. Vol. 60. American Mathematical Society.
[25] Ilan Newman and Christian Sohler. 2013. Every property of hyperfinite graphs is testable. *SIAM J. Comput.* 42, 3 (2013), 1095–1112.

Received 28 February 2021; revised 14 July 2022; accepted 20 September 2022