First Order Definability of Graphs: Tight Bounds on Quantifier Rank *

Oleg Pikhurko
DPMMS, Centre for Mathematical Sciences
Cambridge University, England
O.Pikhurko@dpmms.cam.ac.uk

Helmut Veith
DBAI, Inst. für Informationssysteme
TU Wien, Austria
veith@dbai.tuwien.ac.at

Oleg Verbitsky
Dept. of Mechanics and Mathematics
Kyiv University, Ukraine
oleg@ov.litech.net

Abstract

We investigate upper bounds on the quantifier rank of first order formulas over finite graphs. Our first main result states that any two non-isomorphic graphs G, G' of order n can be distinguished by a first order formula whose quantifier rank is less than or equal $(n+3)/2$. This bound is only $3/2$ greater than a simple lower bound for the number of first order variables necessary to distinguish G and G', which in turn is a lower bound for quantifier rank. As a consequence, we determine, up to an additive constant of 1, the optimum dimension of a variant of the Weisfeiler-Lehman graph canonization algorithm in which the color refinement keeps track of colors of adjoining configurations without their multiplicities.

Our second main result concerns definability of finite graphs. For a graph G of order n, let $D(G)$ denote the minimum quantifier rank of a formula defining G. While it is well-known that $D(G) \leq n+1$ is the best possible general upper bound for $D(G)$, we identify a simple class C of easily recognizable graphs for which $D(G) \leq (n+5)/2$ holds true; for all other graphs, the exact value of $D(G)$ can be easily computed. The defining formulas in this result have only one quantifier alternation and are efficiently constructible. The same result is obtained for the minimal number of variables in a defining first order formula.

1 Introduction

From the logical point of view, a graph G is a structure with a single anti-reflexive and symmetric binary predicate E for the adjacency relation of G. Every closed first order formula Φ with predicate symbols E and $=$ is either true or false on G. Given two non-isomorphic graphs G and G', we say that Φ *distinguishes G from G'* if $G \models \Phi$ but $G' \not\models \Phi$. The *quantifier rank* of Φ is the maximum number of nested quantifiers in this formula (see Section 2 for formal definitions). Let $\mathrm{D}(G, G')$ denote the minimum rank of a formula distinguishing G from G' and $\mathrm{V}(G, G')$ denote the minimum number of variables used in such a formula (different occurrences of the same variable are counted only once). The numbers $\mathrm{D}(G, G')$ and $\mathrm{V}(G, G')$ are symmetric with respect to G and G' because, if Φ distinguishes G from G', then $\neg\Phi$ distinguishes G' from G and has the same quantifier rank and the number of variables. It is not hard to see that $\mathrm{V}(G, G') \leq \mathrm{D}(G, G')$ (see Proposition 1).

First order logic with counting quantifiers is a variant of first order logic. In this logic, a formula $\exists^{\geq m} x\Psi$ expresses that there are at least m distinct vertices x for which Ψ holds. We define $\mathrm{C}(G, G')$ for the logic with counting quantifiers analogously to $\mathrm{V}(G, G')$; it is easy to see that $\mathrm{C}(G, G') \leq \mathrm{V}(G, G')$. $\mathrm{C}(G, G')$ occurs quite naturally in computer science: The distinguishability of graphs with counting quantifiers is closely related to the Weisfeiler-Lehman graph canonization algorithm that was studied since the seventies (see e.g. [3, 7]). It is known that two non-isomorphic graphs G and G' can be recognized to be non-isomorphic by the k-dimensional Weisfeiler-Lehman canonization algorithm for $k = \mathrm{C}(G, G') - 1$ and this takes time $O(k^2 n^{k+1} \log n)$ [7, 13]. On the other hand, if $k < \mathrm{C}(G, G') - 1$, the algorithm fails to distinguish between G and G'. The same correspondence holds between $\mathrm{V}(G, G')$ and the optimum dimension of a simplified variant of the Weisfeiler-Lehman graph canonization algorithm in which the color refinement takes into account only the set of colors of adjoining configurations and ignores their multiplicities [7, Remark 5.5].

Given a graph G, let $\overline{\mathrm{D}}(G)$ be the maximum of $\mathrm{D}(G, G')$ over all non-isomorphic G' of the same order (i.e., number of vertices) as G. Let $\overline{\mathrm{V}}(G)$ and $\overline{\mathrm{C}}(G)$ be defined similarly. The numbers $\overline{\mathrm{D}}(G)$ and $\overline{\mathrm{V}}(G)$ are equal to the minimum quantifier rank and the minimum number of variables suffi-

*This work was supported by the European Community Research Training Network GAMES and the Austrian Science Fund Project Z29-INF. The first author was supported by a research fellowship, St. John's College, Cambridge.

cient to identify, up to isomorphism, a graph G in the class of graphs of the same order. We say that a formula Φ *identifies* a graph G (given the order of G) if Φ distinguishes G from any other non-isomorphic graph G' of the same order. Note that, if G is distinguished from G' by formula $\Phi_{G'}$, then G is identified by the conjunction $\bigwedge_{G'} \Phi_{G'}$ over all non-isomorphic G' of the same order; the quantifier rank of this formula is $\overline{\mathrm{D}}(G)$ and number of variables is $\overline{\mathrm{V}}(G)$. Since in the logic with counting quantifiers graphs of different orders are easily distinguishable, for every G there is a formula with $\overline{\mathrm{C}}(G)$ variables identifying G in the class of *all* graphs, i.e., defining G completely up to isomorphism. Note also that $\overline{\mathrm{C}}(G) - 1$ is equal to the smallest dimension of the Weisfeiler-Lehman algorithm that suffices to obtain a canonic form of G.

It is known that $\overline{\mathrm{C}}(G) = 2$ for almost all graphs of a given order, $\overline{\mathrm{C}}(G) = 3$ for almost all regular graphs of a given order and degree, $\overline{\mathrm{C}}(G) = 2$ for all trees [5, 6, 14, 13], $\overline{\mathrm{C}}(G) = O(g)$ for graphs of genus g [9, 10], and $\overline{\mathrm{C}}(G) \le k + 2$ for graphs of tree-width k [11]. For strongly regular graphs we have $\overline{\mathrm{C}}(G) = O(\sqrt{n}\log n)$, where n denotes the order of G [4]. If G has a separator of size $O(n^{\delta})$, $0 < \delta < 1$, then $\overline{\mathrm{C}}(G) = O(n^{\delta})$ [7]. This result applies, in particular, to classes of graphs with excluded minors that have separators of size $O(\sqrt{n})$ [2].

We define $\mathrm{D}(n)$ (resp. $\mathrm{V}(n)$ and $\mathrm{C}(n)$) to be the maximum of $\overline{\mathrm{D}}(G)$ (resp. $\overline{\mathrm{V}}(G)$ and $\overline{\mathrm{C}}(G)$) over all G of order n. Clearly,

$$\mathrm{C}(n) \le \mathrm{V}(n) \le \mathrm{D}(n). \tag{1}$$

Cai, Fürer, and Immerman [7] prove the remarkable fact that $\mathrm{C}(n) = \Omega(n)$. Though they do not specify the constant hidden in $\Omega(n)$, a simple analysis of their proof gives at least the value $cn/21$, where c is the best possible *expansion* of a 3-regular graph (see [1, Section 3] for the definition of this notion). It is not hard to show that

$$\mathrm{V}(n) \ge n/2 \ \text{ and } \ \mathrm{D}(n) \ge (n+1)/2$$

(see Example 4). On the other hand, an obvious upper bound for the hierarchy (1) is

$$\mathrm{D}(n) \le n. \tag{2}$$

Indeed, a graph G with vertex set $V(G) = \{1, \ldots, n\}$ and edge set $E(G)$ is identified by the formula

$$\exists x_1 \ldots \exists x_n \bigwedge_{\{i,j\} \in E(G)} E(x_i, x_j) \wedge \bigwedge_{\{i,j\} \notin E(G)} \neg E(x_i, x_j).$$

It seems that no better upper bound has been reported in the literature so far. In this paper we obtain a nearly best possible bound.

Theorem 1. $\mathrm{D}(n) \le (n+3)/2$.

It is worth noting that the distinguishing formulas resulting from our proof of Theorem 1 have a rather restricted logical structure. We say that a first order formula Φ is in *negation normal form* if the connective $\neg$ occurs in Φ only in front of atomic subformulas. If Φ is such a formula, its *alternation number* is the maximum number of alternations of $\exists$ and $\forall$ in a sequence of nested quantifiers of Φ. The proof of Theorem 1 produces distinguishing formulas in negation normal form whose alternation number is at most 1.

Our proof of these results is based on the well-known characterization of $\mathrm{D}(G, G')$ and $\mathrm{V}(G, G')$ in terms of the Ehrenfeucht games on G and G' (see Section 2). According to this characterization, $\mathrm{D}(n)$ is the maximum possible length of the game on non-isomorphic graphs of order n, under the condition that the players play optimally. $\mathrm{V}(n)$ is the maximum number of pebbles that may be necessary to finish such a game.

The functions $\mathrm{D}(n)$ and $\mathrm{V}(n)$ are therefore quite meaningful both from the combinatorial point of view and from the point of view of applications to the Weisfeiler-Lehman algorithm. In finite model theory however it is more natural to address defining formulas rather than distinguishing formulas. A first order formula Φ *defines* a graph G if Φ distinguishes G from all non-isomorphic graphs, regardless of their orders. Let $\mathrm{D}(G)$ be the minimum quantifier rank of a formula defining G, and $\mathrm{V}(G)$ be the minimum number of variables in such a formula. We will denote the order of G by n. It is well known that

$$\mathrm{D}(G) \le n + 1$$

for every G (see e.g. [8]). This bound cannot be generally improved because, for example, no formula of quantifier rank n can distinguish between two complete graphs of orders n and $n+1$. Nevertheless, our next aim is to suggest a bound for $\mathrm{D}(G)$ similar to Theorem 1, describe all the exceptions, and compute $\mathrm{D}(G)$ for all exceptional G.

Theorem 2. *There is an efficiently recognizable class of graphs $\mathcal{C}$ such that*

- $\mathrm{D}(G) \le (n+5)/2$ *with the exception of all graphs in $\mathcal{C}$;*
- *if $G \in \mathcal{C}$, then the exact value of* $\mathrm{D}(G)$ *is efficiently computable.*

Moreover, given G, one can efficiently construct its defining formula whose quantifier rank is as small as possible if $G \in \mathcal{C}$ and does not exceed $(n+5)/2$ if $G \notin \mathcal{C}$, and whose alternation number is 1 in both cases.

The same result holds for $\mathrm{V}(G)$. In the above theorem, efficiency means running time $O(n^2 \log n)$ on a random access machine with input graphs given by their adjacency matrices.

The paper is organized as follows. Section 2 contains the relevant definitions from graph theory and logic and the basic facts on the Ehrenfeucht games. In Section 3 we prove Theorems 1 and 2, restated there as Theorems 5 and 7. In Section 4 we extend these results over *colored graphs*. In Section 5 we observe a relation to the complexity-theoretic analysis of the Ehrenfeucht games.

2 Preliminaries

Graphs. Given a graph G, we denote its vertex set by $V(G)$ and its edge set by $E(G)$. The order of G will be sometimes denoted by $|G|$, that is, $|G| = |V(G)|$. The neighborhood of a vertex v consists of all vertices adjacent to v and is denoted by $\Gamma(v)$. We write $G \cong H$ if graphs G and H are isomorphic.

The *complement* of G, denoted by $\overline{G}$, is the graph on the same vertex set $V(G)$ with all those edges that are not in $E(G)$. Given G and G' with disjoint vertex sets, we define the *sum* (or *disjoint union*) $G \sqcup G'$ to be the graph with vertex set $V(G) \cup V(G')$ and edge set $E(G) \cup E(G')$.

A set $S \subseteq V(G)$ is called *independent* (or *stable*) if it contains no adjacent vertices. S is a *clique* if all vertices in S are pairwise adjacent. The *independence* (or *stability*) *number* of G, denoted by $\alpha(G)$, is the largest number of vertices in an independent set of G. The *clique number* of G, denoted by $\omega(G)$, is the biggest number of vertices in a clique of G. The *complete graph* of order n, denoted by K_n, is a graph of order n whose vertex set is a clique. The complement of K_n is the *empty graph* of order n. The *complete bipartite graph* with vertex classes V_1 and V_2, where $V_1 \cap V_2 = \emptyset$, is a graph with vertex set $V_1 \cup V_2$ and edge set consisting of all edges $\{v_1, v_2\}$ for $v_1 \in V_1$ and $v_2 \in V_2$.

If $X \subseteq V(G)$, then $G[X]$ denotes the subgraph induced by G on X. If $X, Y \subseteq V(G)$, then $G[X, Y]$ denotes the bipartite graph induced by G on vertex classes X and Y. A one-to-one map $\phi : S \to S'$, where $S \subseteq V(G)$ and $S' \subseteq V(G')$, is a *partial isomorphism* from G to G' if ϕ is an isomorphism from $G[S]$ to $G'[S']$.

Logic. First order formulas are assumed to be over the set of connectives $\{\neg, \wedge, \vee\}$.

Definition 1. *A* sequence of quantifiers *is a finite word over the alphabet* $\{\exists, \forall\}$. *If S is a set of such sequences, then $\exists S$ (resp. $\forall S$) means the set of concatenations $\exists s$ (resp. $\forall s$) for all $s \in S$. If s is a sequence of quantifiers, then $\bar{s}$ denotes the result of replaing all occurrences of $\exists$ in s by $\forall$, and vice versa. The set $\bar{S}$ consists of all $\bar{s}$ for $s \in S$.*

Given a first order formula Φ, its set of sequences of nested quantifiers *is denoted by* $\mathrm{Nest}(\Phi)$ *and defined by induction as follows:*

1. $\mathrm{Nest}(\Phi) = \{\epsilon\}$ *if Φ is atomic; here, ϵ denotes the empty word.*
2. $\mathrm{Nest}(\neg\Phi) = \overline{\mathrm{Nest}(\Phi)}$.
3. $\mathrm{Nest}(\Phi \wedge \Psi) = \mathrm{Nest}(\Phi \vee \Psi) = \mathrm{Nest}(\Phi) \cup \mathrm{Nest}(\Psi)$.
4. $\mathrm{Nest}(\exists x\Phi) = \exists\,\mathrm{Nest}(\Phi)$ *and* $\mathrm{Nest}(\forall x\Phi) = \forall\,\mathrm{Nest}(\Phi)$.

Definition 2. *The* quantifier rank *of a formula Φ, denoted by $\mathrm{qr}(\Phi)$ is the maximum length of a string in $\mathrm{Nest}(\Phi)$.*

Proposition 1. *Let Φ be a first order formula with $\mathrm{qr}(\Phi) = k$ and suppose that none of variables $x_1, \ldots, x_k$ occurs in Φ. Then there is an equivalent formula Ψ whose bound variables are all in the set $\{x_1, \ldots, x_k\}$.*

We adopt the notion of the *alternation number* of a formula (cf. [15, Definition 2.8]).

Definition 3. *Given a sequence of quantifiers s, let $\mathrm{alt}(s)$ denote the number of occurrences of $\exists\forall$ and $\forall\exists$ in s. The* alternation number *of a first order formula Φ, denoted by $\mathrm{alt}(\Phi)$, is the maximum $\mathrm{alt}(s)$ over $s \in \mathrm{Nest}(\Phi)$.*

Graphs are considered to be structures with a single symmetric and anti-reflexive binary predicate E.

Definition 4. *Given a graph G and a first order formula Φ over vocabulary $\{E, =\}$, we write $G \models \Phi$ if Φ is true on G and $G \not\models \Phi$ otherwise. We say that Φ* distinguishes *G* from *G' if $G \models \Phi$ but $G' \not\models \Phi$. By $\mathrm{D}(G, G')$ (resp. $\mathrm{D}_k(G, G')$) we denote the minimum quantifier rank of a formula (with alternation number at most k resp.) distinguishing G from G'. By $\mathrm{V}(G, G')$ we denote the minimum l such that over the variable set $\{x_1, \ldots, x_l\}$ there is a formula distinguishing G from G'.*

Note that

$$\mathrm{V}(G, G') \le \mathrm{D}(G, G') \le \mathrm{D}_k(G, G') \le \mathrm{D}_{k-1}(G, G')$$

for every $k \ge 1$, where the first inequality follows from Proposition 1.

Definition 5. *We say that Φ* defines *G (up to isomorphism) if Φ distinguishes G from any non-isomorphic graph G'. By $\mathrm{D}(G)$ (resp. $\mathrm{D}_k(G)$) we denote the minimum quantifier rank of a formula defining G (with alternation number at most k resp.). By $\mathrm{V}(G)$ we denote the minimum l such that over the variable set $\{x_1, \ldots, x_l\}$ there is a formula defining G.*

Proposition 2.

$$\begin{aligned}
\mathrm{D}(G) &= \max\{\,\mathrm{D}(G, G') : G' \not\cong G\}\,, \\
\mathrm{D}_k(G) &= \max\{\,\mathrm{D}_k(G, G') : G' \not\cong G\}\,, \\
\mathrm{V}(G) &= \max\{\,\mathrm{V}(G, G') : G' \not\cong G\}\,.
\end{aligned}$$

Proof. We prove the first equality; the proof of the other two is similar. Given G' non-isomorphic with G, let $\Phi_{G'}$ be a formula of minimum quantifier rank distinguishing G from G', that is, $\mathrm{qr}(\Phi_{G'}) = \mathrm{D}(G, G')$. Let $R = \max_{G'} \mathrm{qr}(\Phi_{G'})$. We have $\mathrm{D}(G) \geq R$ because $\mathrm{D}(G) \geq \mathrm{D}(G, G')$ for every G'. To prove the reverse inequality $\mathrm{D}(G) \leq R$, notice that G is defined by the formula $\Phi = \bigwedge_{G'} \Phi_{G'}$ whose quantifier rank is R. The only problem is that Φ is an infinite conjunction (i.e., a $FO_{\infty\omega}$-formula.) However, as is well known, over a fixed finite vocabulary there are only finitely many inequivalent first order formulas of bounded quantifier rank (see e.g. [7, 8, 12]). We therefore can reduce Φ to a finite conjunction. □

Proposition 3.

1. *$\mathrm{D}(G, G') = \mathrm{D}(\overline{G}, \overline{G'})$, $\mathrm{D}_k(G, G') = \mathrm{D}_k(\overline{G}, \overline{G'})$, and $\mathrm{V}(G, G') = \mathrm{V}(\overline{G}, \overline{G'})$.*

2. *$\mathrm{D}(G) = \mathrm{D}(\overline{G})$, $\mathrm{D}_k(G) = \mathrm{D}_k(\overline{G})$, and $\mathrm{V}(G) = \mathrm{V}(\overline{G})$.*

We omit the simple proof of this proposition.

Games. The *Ehrenfeucht game* is played on a pair of structures of the same vocabulary. To simplify the exposition, our definition is restricted to graphs.

Definition 6. *Let G and G' be graphs with disjoint vertex sets. The r-round l-pebble Ehrenfeucht game on G and G', denoted by $\mathrm{EHR}_r^l(G, G')$, is played by two players, Spoiler and Duplicator, using l pairwise distinct pebbles $p_1, \ldots, p_l$, each given in duplicate. Spoiler starts the game. A* round *consists of a move of Spoiler followed by a move of Duplicator. At each move Spoiler takes a pebble, say p_i, selects one of the graphs G or G', and places p_i on a vertex of this graph. In response Duplicator should place the other copy of p_i on a vertex of the other graph. It is allowed to move previously placed pebbles to another vertex and place more than one pebble on the same vertex.*

After each round of the game, for $1 \leq i \leq l$ let x_i (resp. y_i) denote the vertex of G (resp. G') occupied by p_i, irrespectively of who of the players placed the pebble on this vertex. If p_i is off the board at this moment, x_i and y_i are undefined. If after each of r rounds it is true that

$$x_i = x_j \text{ iff } y_i = y_j \text{ for all } 1 \leq i < j \leq l,$$

and the component-wise correspondence $(x_1, \ldots, x_l)$ to $(y_1, \ldots, y_l)$ is a partial isomorphism from G to G', this is a win for Duplicator; otherwise the winner is Spoiler.

*The k-*alternation *Ehrenfeucht game on G and G' is a variant of the game in which Spoiler is allowed to switch from one graph to another at most k times during the game, i.e., in at most k rounds he can choose the graph other than that in the preceding round.*

The main technical tool we will use is given by the following statement.

Proposition 4.

1. *$\mathrm{V}(G, G')$ equals the minimum l such that Spoiler has a winning strategy in $\mathrm{EHR}_r^l(G, G')$ for some r.*

2. *$\mathrm{D}(G, G')$ equals the minimum r such that Spoiler has a winning strategy in $\mathrm{EHR}_r^r(G, G')$.*

3. *$\mathrm{D}_k(G, G')$ equals the minimum r such that Spoiler has a winning strategy in the k-alternation $\mathrm{EHR}_r^r(G, G')$.*

We refer the reader to [12, Theorem 6.10] for the proof of the first claim, to [8, Theorem 1.2.8], [12, Theorem 6.10], or [17, Theorem 2.3.1] for the second claim, and to [16] for the third claim. Note that in the first claim one can put the bound $r \leq n^l$, that is, if Spoiler is able to win with l pebbles, he needs not too many rounds.

Proposition 4 immediately implies Propositions 1 and 3.

Remark 3. *If we prohibit removing pebbles from one vertex to another in $\mathrm{EHR}_r^r(G, G')$, this will not affect the outcome of the game.*

Definition 7. *We denote the variant of $\mathrm{EHR}_r^r(G, G')$ with removing pebbles prohibited by $\mathrm{EHR}_r(G, G')$.*

The examples below are obtained by simple application of Proposition 4.

Example 4.

1. $\mathrm{V}(K_m \sqcup \overline{K_m}, K_{m+1} \sqcup \overline{K_{m-1}}) = m,$
   $\mathrm{D}(K_m \sqcup \overline{K_m}, K_{m+1} \sqcup \overline{K_{m-1}}) = m + 1.$

2. $\mathrm{V}(K_{m+1} \sqcup \overline{K_m}, K_m \sqcup \overline{K_{m+1}}) =$
   $\mathrm{D}(K_{m+1} \sqcup \overline{K_m}, K_m \sqcup \overline{K_{m+1}}) = m + 1.$

3 Main results

Definition 8. *We call vertices u and v of a graph G* similar *and write $u \sim v$ if every third vertex t is simultaneously adjacent or not to u and v. Let $[u]_G = \{ v \in V(G) : v \sim u \}$, $\sigma_G(u) = |[u]_G|$, and $\sigma(G) = \max_{u \in V(G)} \sigma_G(u)$. If the graph is clear from the context, the subscript G may be omitted.*

Proposition 5.

1. *$\sim$ is an equivalence relation on $V(G)$.*

2. *Every equivalence class $[u]$ is either a clique or an independent set.*

Proof. The proposition is straightforward. The only care should be taken to check transitivity: if $u \sim v$ and $v \sim w$, then $u \sim w$. For every $t \neq u, w$, we need to show that u and t are adjacent iff so are w and t. If $t \neq v$, this is true because both adjacencies are equivalent to the adjacency of v and t. There remains the case that $t = v$. Then u and v are adjacent iff so are u and w (as $v \sim w$), which in turn holds iff so are w and v (as $u \sim v$). □

Below, referring to efficient algorithms, we mean random access machines whose running time on graphs of order n, represented by adjacency matrices, is $O(n^2 \log n)$.

Proposition 6.

1. *There is an efficient algorithm that, given G, finds the partition of $V(G)$ into classes of pairwise similar vertices.*
2. *Given G, the number $\sigma(G)$ is efficiently computable.*

Proof. Notice that non-adjacent vertices are in the same similarity class iff the corresponding rows of the adjacency matrix are identical. Thus, in order to find similarity classes containing more than one element, it suffices, using the standard $O(n \log n)$-comparison sorting, to arrange the rows of the adjacency matrices of G and $\overline{G}$ in lexicographic order. □

Definition 9. *If $v \in V(G)$, the notation $H = G \oplus v$ means that H is a graph obtained from G by adding a new vertex v' so that $[v]_{G'} = [v]_G \cup \{v'\}$. In other words, v' is similar to v and adjacent to v depending on if $[v]_G$ is a clique or an independent set. If $\sigma_G(v) = 1$, then H is not uniquely defined, and can obtain two possible forms depending on whether v and v' are adjacent. Furthermore, we define $G \oplus 0v = G$ and $G \oplus kv = (G \oplus (k-1)v) \oplus v$ for a positive integer k in the natural way.*

Convention. In the sequel, writing $H = G \oplus kv$ we will tacitly assume that H is an arbitrary isomorphic copy of $G \oplus kv$. When considering the Ehrenfeucht game on G and H, we will in addition suppose that the vertex sets of these graphs are disjoint.

Lemma 1. *If G and G' are non-isomorphic graphs of orders n and n' respectively and $n \leq n'$, then*

$$\mathrm{D}_1(G, G') \leq (n+4)/2 \tag{3}$$

unless $G' = G \oplus (n'-n)v$ for some $v \in V(G)$.

Proof. We will describe a strategy of Spoiler winning $\textsc{Ehr}_r(G, G')$ for $r = \lfloor (n+4)/2 \rfloor$ unless $G' = G \oplus (n'-n)v$. The strategy splits the game in two phases. During Phase 1 Spoiler will select vertices of G only. Let $x_i \in V(G)$ denote the vertex selected by Spoiler in the i-th round of Phase 1 and $X_i = \{x_1, \ldots, x_i\}$. Given two vertices u and v in the complement $\overline{X_i} = V(G) \setminus X_i$, we call them X_i*-similar* if $\Gamma(u) \cap X_i = \Gamma(v) \cap X_i$. Obviously, X_i-similar vertices are also X_{i-1}-similar, and X_i-similarity is an equivalence relation on $\overline{X_i}$. The strategy of Spoiler in Phase 1 is to select a vertex x_{i+1} so that the number of equivalence classes strictly increases. When no such vertex exists, Phase 1 ends.

Suppose that Phase 1 lasted for s rounds and let t be the number of equivalence classes with respect to X_s-similarity. As initially $X_0 = \emptyset$ and there was only one equivalence class, we have $t \geq s+1$ and therefore $s \leq (n-1)/2$.

Denote the t equivalence classes by $C_1, \ldots, C_t$. Let us explore the structure of the partition $\overline{X_s} = C_1 \cup \ldots \cup C_t$.

Claim A. Every C_i is either a clique or an independent set.

Proof of Claim. Suppose not. Then C_i contains three vertices u, v, and w such that u and v are adjacent but u and w are not. However, if we move u to X_s, then C_i splits into two classes, which are non-empty as they contain u and w respectively. Hence the number of equivalence classes increases at least by 1, a contradiction. □

Claim B. If C_i and C_j contain at least 2 vertices each, then the bipartite graph $G[C_i, C_j]$ is either complete or empty.

Proof of Claim. Suppose not: for example, $u \in C_i$ is adjacent to $v \in C_j$ but not to $w \in C_j$. If we move u to X_s, then C_j splits into two non-empty classes and $C_i \setminus \{u\}$ stays non-empty. Again the number of equivalence classes increases, a contradiction. □

Let $y_i \in V(G')$ denote the vertex selected by Duplicator in the i-th round of Phase 1 and $C'_1, \ldots, C'_{t'}$ be the equivalence classes in G' with respect to $\{y_1, \ldots, y_s\}$-similarity. Define $N(C_i)$ for $i \leq t$ to be the set of indices m such that each vertex of C_i is adjacent to x_m. Define $N(C'_j)$ for $j \leq t'$ similarly. Clearly, all $N(C_i)$ and all $N(C'_j)$ are pairwise distinct. Note also that every $N(C_i)$ should be equal to one of the $N(C'_j)$'s and vice versa for otherwise Spoiler wins in the next move. We can therefore assume that $t' = t$ and $N(C_i) = N(C'_i)$ for every i.

We now describe Spoiler's strategy in Phase 2 of the game. Without loss of generality, assume that $|C_i| \geq 2$ for $i \leq q$ and $|C_i| = 1$ for $i \geq q+1$. It is easy to see that either Spoiler wins in 2 extra moves (altogether $s + 2 \leq (n-1)/2 + 2 = (n+3)/2$ moves), possibly alternating once between G and G', or the following conditions are met:

- $|C'_i| = 1$ iff $i \geq q+1$.
- Each $G'[C'_i]$ or $G'[C'_i, C'_j]$ for $i, j \leq q$ is either empty or complete.

- If $i \le q$, then $G'[C'_i]$ is complete iff $G[C_i]$ is.
- If $i, j \le q$, then $G'[C'_i, C'_j]$ is complete iff $G[C_i, C_j]$ is.

We call C_i *useful* if $|C_i| \ne |C'_i|$. If C_i is useful, then Spoiler is able to win having made in Phase 2 at most $|C_i| + 1$ moves and at most 1 alternation between G and G'. To do so, Spoiler selects $\min\{|C_i|, |C'_i|\} + 1$ vertices in the larger of the classes C_i and C'_i.

Suppose that there are two useful classes, C_i and C_j. Observe that

$$|C_i|+|C_j| = |\overline{X_s}| - \sum_{l \ne i,j} |C_l| \le (n-s)-(t-2) \le n-2s+1.$$

It follows that one of the useful classes has at most $(n - 2s + 1)/2$ vertices and Spoiler wins altogether in at most $s + (n - 2s + 1)/2 + 1 = (n + 3)/2$ moves with at most 1 alternation. In the rest of the proof we therefore consider the case that there is at most one useful class, that is, $|C_i| = |C'_i|$ for all i, with at most one exception.

Denote $Z = \bigcup_{m=q+1}^{t} C_m$ and $Z' = \bigcup_{m=q+1}^{t} C'_m$. Given $i \le t$ and $A \subseteq \{q + 1, \ldots, t\}$, we define $K_{i,A} = \left\{u \in C_i : \Gamma(u) \cap Z = \bigcup_{m \in A} C_m\right\}$. A counterpart of $K_{i,A}$ in G' is defined by $K'_{i,A} = \left\{u \in C'_i : \Gamma(u) \cap Z' = \bigcup_{m \in A} C'_m\right\}$. Note that all non-empty $K_{i,A}$ (resp. $K'_{i,A}$) form the partition of $\overline{X_s} \subset V(G)$ (resp. $\overline{X'_s} \subset V(G'))$ into the classes of pairwise similar vertices as defined in Definition 8.

We call $K_{i,A}$ *useful* if $|K_{i,A}| \ne |K'_{i,A}|$. If $K_{i,A}$ is useful, then Spoiler is able to win having made in Phase 2 at most $|K_{i,A}|+2$ moves (actually at most $\min\{|K_{i,A}|, |K'_{i,A}|\}+2$) and at most 1 alternation between G and G'. If, for example, $|K_{i,A}| < |K'_{i,A}|$, then Spoiler selects $|K_{i,A}| + 1$ vertices in $K'_{i,A}$. Duplicator is forced to select at least one vertex $x \notin K_{i,A}$. If $x \notin C_i$, this is an immediate win of Spoiler. If $x \in C_i$, then there is $m > q$ such that exactly one of the two adjacencies takes place: either x is adjacent to the vertex in C_m or y is adjacent to the vertex in C'_m, where $y \in K'_{i,A}$ is the vertex selected by Spoiler in the same round when x is selected by Duplicator. In the next move Spoiler selects the vertex in C'_m and wins.

Assume that there are two useful classes, $K_{i,A}$ and $K_{i,B}$. Observe that

$$\begin{aligned} |K_{i,A}| + |K_{i,B}| \le |C_i| &= |\overline{X_s}| - \sum_{l \ne i} |C_l| \\ &\le (n-s) - (t-1) \le n - 2s. \end{aligned}$$

It follows that one of the useful classes has at most $(n - 2s)/2$ vertices and Spoiler wins altogether in at most $s + (n-2s)/2+2 = (n+4)/2$ moves with at most 1 alternation.

Note that the partial isomorphism established by Duplicator in Phase 1 extends to the isomorphism between G and G' iff $|K_{i,A}| = |K'_{i,A}|$ for all i and A. Since G and G' are non-isomorphic, there is at least one useful $K_{i,A}$. If for this i it holds $|C_i| = |C'_i|$, then there must be another useful $K_{i,B}$ and therefore our strategy for Spoiler ensures the bound (3). If there are two useful K_{i_1,A_1} and K_{i_2,A_2} with $i_1 \ne i_2$, then at least one of the equalities $|C_{i_1}| = |C'_{i_1}|$ or $|C_{i_2}| = |C'_{i_2}|$ holds and we again have (3). Thus, this bound may be false only if there is a unique useful $K_{i,A}$, that is, $n' > n$, $|K'_{i,A}| = |K_{i,A}|+(|C'_i|-|C_i|) = |K_{i,A}|+(n'-n)$, and $|K_{j,B}| = |K'_{j,B}|$ for all other pairs $(j, B) \ne (i, A)$. Assume that these conditions are met. Another condition we assume is $K_{i,A} \ne \emptyset$, because otherwise Phase 2 lasts at most 2 rounds and (3) is true.

It remains to notice that, if we remove $n' - n$ vertices from $K'_{i,A}$, we obtain a graph isomorphic to G. It follows that $G' = G \oplus (n' - n)v$ with $v \in K_{i,A}$. □

We now restate and prove Theorem 1 from the introduction.

Theorem 5. *If G and G' are non-isomorphic and have the same order n, then*

$$(n + 1)/2 \le \mathrm{D}(G, G') \le \mathrm{D}_1(G, G') \le (n + 3)/2. \quad (4)$$

Proof. The lower bound is due to Example 4. Lemma 1 immediately gives an upper bound of $(n+4)/2$, which is a bit worse than we now claim. To improve it, we go trough the lines of the proof of Lemma 1 but make use of the equality $n = n'$. The latter causes the following changes. The bound (4) might be now wrong only if $|C_i| = |C'_i|$ for all i. This condition, together with the non-isomorphism of G and G', implies that, for some i, there are two useful classes $K_{i,A}$ and $K_{i,B}$ with $|K_{i,A}| < |K'_{i,A}|$ and $|K_{i,B}| > |K'_{i,B}|$. We have

$$\begin{aligned} 2|K_{i,A}| &+ 2|K'_{i,B}| + 2 \\ &\le |K_{i,A}| + |K_{i,B}| + |K'_{i,A}| + |K'_{i,B}| \\ &\le |C_i| + |C'_i| = 2|C_i|. \end{aligned}$$

It follows that at least one of $|K_{i,A}|$ or $|K'_{i,B}|$ does not exceed $(|C_i| - 1)/2$ and this gives us the desired gain of $1/2$ in the bound. □

Remark 6. *Theorem 5 states the existence of a formula with quantifier rank $(n + 3)/2$ distinguishing two given graphs of order n. The proofs of Lemma 1 and Theorem 5 can actually give more: Given a graph G, in time $O(n^2 \log n)$ on a random access machine one can construct a formula $\Phi(G)$ with $\mathrm{qr}(\Phi(G)) \le (n + 3)/2$ distinguishing G from* every *non-isomorphic graph G' of the same order n. $\Phi(G)$ is a kind of the* Hintikka formula *of G [8, Definition 1.2.5] designed according to the particular Spoiler's strategy.*

Note that the efficient constructibility of $\Phi(G)$ reduces the instance of the graph isomorphism problem whether or not $G \cong G'$ to the instance of the model checking problem *whether or not $G' \models \Phi(G)$. The latter can be solved in time $n^{\mathrm{qr}(\Phi(G))+O(1)}$ on a random access machine, which is essentially no worse than the time bound for the variant of the Weisfeiler-Lehman algorithm discussed in Sections 1 and 4.*

Lemma 2. *Let x_i (resp. y_i) denote the vertex of G (resp. G') selected in the i-th round of $\mathrm{EHR}_r(G, G')$. Then, as soon as a move of Duplicator violates the condition that $x_i \sim x_j$ iff $y_i \sim y_j$, Spoiler wins either immediately or in the next move with alternation between the graphs.*

Proof. Suppose that Duplicator selects y_j so that $y_i \not\sim y_j$ while $x_i \sim x_j$ for some $i < j$ and the correspondence between the x_m's and the y_m's, $1 \le m \le j$, is still a partial isomorphism. Then there is $y \in V(G')$ adjacent to exactly one of y_i and y_j. Note that such a y could not be selected by the players previously. In the next move Spoiler selects y and wins, whatever the move of Duplicator. $\square$

Lemma 3. *Let $|G| = n$, $v \in V(G)$, $\sigma_G(v) = s$, and $G' = G \oplus kv$ with $k \ge 1$. Then*

$$s+1 \le \mathrm{V}(G,G') \le \mathrm{D}_1(G,G') \le s+1+\frac{n+1}{s+1}$$
$$\le \begin{cases} (n+5)/2 & \text{for } 1 \le s \le (n-1)/2 \\ s+3-1/(n/2+1) & \text{for } s \ge n/2. \end{cases}$$

Proof. The lower bound is given by the following strategy for Duplicator in $\mathrm{EHR}_r^s(G, G')$. Whenever Spoiler selects a vertex outside $[v]$ in either graph, Duplicator selects its copy in the other graph. If Spoiler selects an unoccupied vertex similar to v, then Duplicator selects an arbitrary unoccupied vertex similar to v in the other graph. Clearly, this strategy preserves the isomorphism arbitrarily long, that is, is winning for every r.

The upper bound for $\mathrm{D}_1(G, G')$ is ensured by the following Spoiler's strategy winning in the 1-alternation $\mathrm{EHR}_r(G, G')$ for $r = \lfloor s + 1 + \frac{n+1}{s+1} \rfloor$. In the first round Spoiler selects a vertex in $[v]_{G'}$. Suppose that Duplicator replies with a vertex in $[u]_G$.

Case 1: $|[u]_G| \le s$. Spoiler continues to select vertices in $[v]_{G'}$. In the $(s+1)$-st round at latest, Duplicator selects a vertex outside $[u]_G$. Spoiler wins in the next move by Lemma 2, having made at most $s+2$ moves and one alternation.

Case 2: $|[u]_G| \ge s+1$. Spoiler selects one vertex in each similarity class of G' containing at least $s+1$ vertices. Besides $[v]_{G'}$, there can be at most $\frac{n-s}{s+1}$ such classes. At latest in the $\lfloor \frac{n-s}{s+1} + 1 \rfloor$-th round Duplicator selects either another vertex in a class with an already selected vertex (then Spoiler wins in one extra move by Lemma 2) or a vertex in $[w]_G$ with $|[w]_G| \le s$. In the latter case Spoiler selects s more vertices in the corresponding class of G'. Duplicator is forced to move outside $[w]_G$ and loses in the next move by Lemma 2. Altogether there are made at most $\lfloor \frac{n-s}{s+1} + 1 \rfloor + s + 1 \le s + 1 + \frac{n+1}{s+1}$ moves.

If $s \ge n/2$, the last inequality of the lemma is straightforward and, if $1 \le s \le (n-1)/2$, it follows from the fact that the function $f(x) = x + \frac{n+1}{x}$ on $[2, (n+1)/2]$ attains its maximum at the endpoints of this range. $\square$

Using Lemma 3, Lemma 1 can now be refined (at the cost of weakening the bound by $1/2$).

Lemma 4. *If G and G' are graphs of orders $n \le n'$, then*

$$\mathrm{D}_1(G,G') \le (n+5)/2$$

unless

$$\sigma(G) \ge n/2 \text{ and } G' = G \oplus (n'-n)v \text{ for some } v \in V(G) \text{ with } \sigma_G(v) = \sigma(G). \tag{5}$$

In the latter case we have

$$\sigma(G)+1 \le \mathrm{V}(G,G') \le \mathrm{D}_1(G,G') \le \sigma(G)+2. \tag{6}$$

Note that the condition (5) determines G' up to isomorphism with two exceptions if n is even. Namely, for $G = K_m \sqcup \overline{K_m}$ and $G = \overline{K_m \sqcup \overline{K_m}}$ there are two ways to extend G to G'.

The gap between the bounds (6) can be completely closed.

Lemma 5. *Let G and G' be graphs of orders $n \le n'$, and assume condition (5). Then $\mathrm{V}(G, G') = \sigma(G) + 1$ for all such G and G', $\mathrm{D}_0(G, G') = \sigma(G)+1$ if $[v]_G$ is a maximal clique or independent set, and $\mathrm{D}(G, G') = \sigma(G)+2$ if $[v]_G$ is not.*

Using the inequality $|[v]_G| \ge n/2$ and the mutual similarity of vertices in $[v]_G$, one can easily show that $[v]_G$ is a maximal (with respect to inclusion) clique or independent set iff $|[v]_G| = \alpha(G)$ or $|[v]_G| = \omega(G)$ respectively. However, the former condition is more preferable because it is efficiently verifiable.

Proof. For simplicity we assume tha t $[v]_G$ is an independent set. Otherwise we can switch to $\overline{G}$ and $\overline{G'}$ by Proposition 3. Denote $s = \sigma(G) = |[v]_G|$.

Case 1: $[v]_G$ is maximal independent. We show the bound $\mathrm{D}_0(G, G') \le s+1$ by describing Spoiler's strategy winning the 0-alternation $\mathrm{EHR}_{s+1}(G, G')$. Spoiler selects $s+1$ vertices in $[v]_{G'}$. Duplicator is forced to select at least one vertex $u_1 \in [v]_G$ and at least one vertex $u_2 \notin [v]_G$. Since $[v]_G$ is a maximal independent set, u_1 and u_2 are adjacent and this is Spoiler's win.

Case 2: $[v]_G$ is not maximal. We first show the bound $\mathrm{V}(G, G') \le s+1$ by describing Spoiler's strategy winning $\mathrm{EHR}_{s+2}^{s+1}(G, G')$. As in the preceding case, Spoiler selects $s+1$ vertices in $[v]_{G'}$ and there are $u_1 \in [v]_G$ and

$u_2 \notin [v]_G$ selected in response by Duplicator. Assume that u_1 and u_2 are not adjacent for otherwise Duplicator loses immediately. Since u_1 and u_2 are not similar, there is $u \in V(G) \setminus \{u_1, u_2\}$ adjacent to exactly one of u_1 and u_2. It follows that $u \notin [v]_G$. Note that u could not be selected by Duplicator in the first $s+1$ rounds without immediately losing. Therefore, Duplicator has selected in $[v]_G$ at least two vertices, say, u_0 and u_1. In the $(s+2)$-nd round Spoiler moves the pebble from u_0 to u and wins because the counterparts of u_1 and u_2 in G' are similar and hence equally adjacent or non-adjacent to any counterpart of u.

We now show the bound $\mathrm{D}(G, G') > s+1$ by describing Duplicator's strategy winning $\mathrm{E{\scriptsize HR}}_{s+1}(G, G')$. Whenever Spoiler selects a vertex of either graph, Duplicator selects its copy in the other graph, with the convention that the copy of a vertex in $[v]_{G'}$ is an arbitrary unselected vertex in $[v]_G$. This is impossible only in the case when Spoiler selects $s+1$ vertices all in $[v]_{G'}$. Then Duplicator, in addition to s vertices of $[v]_G$, selects one more vertex extending $[v]_G$ to a larger independent set. $\square$

Definition 10. *$\mathcal{S}$ is the class of graphs G with $\sigma(G) > (|G|+3)/2$. $\mathcal{S}_1$ is the class of graphs G with $\sigma(G) > (|G|+3)/2$ such that the largest similarity class is a maximum independent set or clique. $\mathcal{S}_2$ is the class of graphs G with $\sigma(G) > (|G|+1)/2$ such that the largest similarity class is not a maximum independent set or clique.*

Theorem 7. $\mathrm{V}(G) \le (|G|+5)/2$ *with the exception of all graphs in $\mathcal{S}$. If $G \in \mathcal{S}$, then* $\mathrm{V}(G) = \sigma(G)+1$.

$\mathrm{D}_1(G) \le (|G|+5)/2$ *with the exception of all graphs in $\mathcal{S}_1 \cup \mathcal{S}_2$. If $G \in \mathcal{S}_1$, then* $\mathrm{D}(G) = \sigma(G)+1$*; If $G \in \mathcal{S}_2$, then* $\mathrm{D}(G) = \sigma(G)+2$.

Proof. We prove the theorem for $\mathrm{V}(G)$; the proof for $\mathrm{D}(G)$ is completely similar. Recall that, by Proposition 2, $\mathrm{V}(G) = \max\{\mathrm{V}(G, G') : G' \not\cong G\}$. We consider two cases.

Case 1: $\sigma(G) < |G|/2$. For every $G' \not\cong G$ we have $\mathrm{V}(G, G') \le (|G|+5)/2$ by Lemma 4. Since $G \notin \mathcal{S}$, the theorem in Case 1 is true.

Case 2: $\sigma(G) \ge |G|/2$. If $G' = G \oplus kv$ for some $k \ge 1$ and $v \in V(G)$ such that $\sigma_G(v) = \sigma(G)$, then $\mathrm{V}(G, G') = \sigma(G)+1$ by Lemma 5. By the definition of $\mathcal{S}$, we therefore have $\mathrm{V}(G, G') \le (|G|+5)/2$ if $G \notin \mathcal{S}$ and $\mathrm{V}(G, G') > (|G|+5)/2$ if $G \in \mathcal{S}$.

If $G = G' \oplus kv$ for some $k \ge 1$, G' such that $\sigma(G') \ge |G'|/2$, and $v \in V(G')$ such that $\sigma_{G'}(v) = \sigma(G')$, then $\mathrm{V}(G, G') = \sigma(G')+1$ by Lemma 5 and hence $\mathrm{V}(G, G') < \sigma(G)+1$.

If G' is any other graph non-isomorphic with G, then $\mathrm{V}(G, G') \le (|G|+5)/2$ by Lemma 4.

Summarizing, if $G \notin \mathcal{S}$, we have $\max_{G'} \mathrm{V}(G, G') \le (|G|+5)/2$ and, if $G \in \mathcal{S}$, we have $\max_{G'} \mathrm{V}(G, G') = \sigma(G)+1 > (|G|+5)/2$. Thus, in Case 2 the theorem is also true. $\square$

A variant of Theorem 7 was stated in the introduction as Theorem 2. The efficiency statements of Theorem 2 follow from Proposition 6. The efficient constructibility of a defining formula follows from the fact that the descriptions of Spoiler's strategies in the proofs of Lemmas 1, 3, and 5 can be converted into an algorithm that, given G, in time $O(n^2 \log n)$ constructs its defining formula with alternation number 1 and quantifier rank precisely $\mathrm{D}(G)$ if $G \in \mathcal{S}_1 \cup \mathcal{S}_2$ and at most $(|G|+5)/2$ otherwise (see, in particular, Remark 6).

4 Extension to colored graphs

A *colored graph* is a structure that, in addition to the anti-reflexive and symmetric binary relation, has countably many unary relations C_i, $i \ge 1$. The truth of $C_i(v)$ for a vertex v is interpreted as coloration of v in color i. An isomorphism of colored graphs preserves the adjacency relation and, moreover, matches a vertex of one graph to an equally colored vertex of the other graph. We consider finite colored graphs, whose vertices can have only finitely many colors.

Theorem 1 holds true for colored graphs literally. If we try to state Theorem 2 for colored graphs, we face an obstacle: The notion of a defining formula becomes questionable. For example, the simplest uncolored one-vertex graph cannot be distinguished from all its colorations by one and the same first order formula. We therefore should allow defining formulas in the infinitary logic $FO_{\infty\omega}$, in particular, keeping for colored graphs the fact that $\mathrm{D}(G)$ is the maximum of $\mathrm{D}(G, G')$ over all G' non-isomorphic with G. Then Theorem 2 carries over to colored graphs as well. We now list a few, more or less evident, changes that are necessary in the proofs of Theorems 1 and 2.

First of all, we modify the notion of vertex similarity adding the requirement that similar vertices are equally colored. Definitions 8 and 9 now have perfect sense for colored graphs. In the proof of Lemma 1 we also add the requirement that X_i-similar vertices are equally colored. In Lemma 5 and Definition 10, a maximal independent set or clique now becomes a maximum *monochromatic* independent set or clique. With these modifications, all the proofs remain virtually the same.

Cai, Fürer, and Immerman [7] relate the number of variables in a formula distinguishing two (colored) graphs with the dimension of a variant of the Weisfeiler-Lehman algorithm recognizing non-isomorphism of these graphs in which each color refinement round records only one of equally colored adjoining configurations. The authors show that the smallest dimension that suffices to find a canonic form of every graph on n vertices is equal to $\mathrm{V}(n)-1$ [7, Remark 5.5]. Theorem 1 together with the lower bound of Example 4 implies therefore the following fact.

Corollary 1. *The optimum dimension of the multiplicity-ignoring variant of the Weisfeiler-Lehman algorithm on graphs of order n is at least $n/2-1$ and at most $n/2+1/2$, that is, is equal either to $\lceil n/2-1\rceil$ or to $\lfloor n/2+1/2\rfloor$.*

In the context of the Weisfeiler-Lehman algorithm, Immerman and Lander [13] associate with every class of graphs $\mathcal{X}$ the function Var $(\mathcal{X}, n)$ that is defined as follows. Given a colored graph K, its vertex u, and a first order formula Φ with at most one free variable, we write $K, u \models \Phi$ if Φ is true on K with the value of the free variable assigned to u. Given two colored graphs K, H and their vertices $u \in V(K)$, $v \in V(H)$, we write $(K,u) \cong (H,v)$ if there is an isomorphism from K to H taking u to v. We say that a formula Φ with at most one free variable *distinguishes* (K,u) from (H,v) if $K,u \models \Phi$ but $H,v \not\models \Phi$. If $(K,u) \not\cong (H,v)$, let $\mathrm{V}(K,u,H,v)$ denote the minimum number of variables in a formula distinguishing (K,u) from (H,v). Given a graph G, let $\hat{\mathrm{V}}(G)$ be the maximum of $\mathrm{V}(\hat{G},u,H,v)$ over all colorations $\hat{G}$ of G, colored graphs H, and vertices $u \in V(\hat{G})$, $v \in V(H)$. Then Var $(\mathcal{X}, n)$ is the maximum of $\hat{\mathrm{V}}(G)$ over all G in $\mathcal{X}$ of order at most n.

For the class $\mathcal{G}$ of all graphs, Immerman and Lander prove that Var $(\mathcal{G}, n) = n+1$ [13, Proposition 1.4.3]. Our Theorem 7, generalized to colored graphs, improves this bound in the following way.

Corollary 2. *There is a class $\mathcal{S}$ of graphs of simple, easily recognizable, structure such that*

$$\text{Var}\,(\mathcal{G}\setminus\mathcal{S}, n) \le (n+5)/2. \tag{7}$$

On the other hand, if $G \in \mathcal{S}$, then $\hat{\mathrm{V}}(G) > (|G|+5)/2$ but the exact value of $\hat{\mathrm{V}}(G)$ is easy to compute.

More specifically, $\mathcal{S}$ is as defined in Definition 10 and

$$\hat{\mathrm{V}}(G) = \sigma(G)+1 \text{ for all } G \in \mathcal{S}. \tag{8}$$

Proof. If $K \not\cong H$, then

$$\max_{u,v} \mathrm{V}(K,u,H,v) = \mathrm{V}(K,H).$$

Indeed, the inequality $\mathrm{V}(K,u,H,v) \le \mathrm{V}(K,H)$ for all u, v is straightforward. To show that for some u and v we have the equality, let $\Phi_{u,v}(x)$ be a formula with the smallest number of variables distinguishing (K,u) from (H,v). We can assume that each $\Phi_{u,v}(x)$ really contains the free variable x for else a closed $\Phi_{u,v}$ distinguishes K and H, and we are done immediately. Then, for an arbitrary $u \in V(K)$, the formula $\exists x \bigwedge_{v\in V(H)} \Phi_{u,v}(x)$ distinguishes G from H and has as many variables as some $\Phi_{u,v}(x)$ has.

If $K \cong H$ and ϕ is an isomorphism from H to K, then obviously $\mathrm{V}(K,u,H,v) = \mathrm{V}(K,u,\phi(H),\phi(v))$ and hence in this case

$$\max_{u,v} \mathrm{V}(K,u,H,v) = \max_{u,v} \mathrm{V}(K,u,K,v),$$

where the latter maximum is over vertices u and v of K that cannot be taken to one another by any automorphism of K, that is, are in different orbits of the group of automorphisms $\mathrm{Aut}(G)$.

Summarizing, we conclude that

$$\hat{\mathrm{V}}(G) = \max_{\hat{G}}\{\mathrm{V}(\hat{G}), \max_{u,v} \mathrm{V}(\hat{G},u,\hat{G},v)\}, \tag{9}$$

where $\hat{G}$ is a coloration of G, and u and v are in different orbits of $\mathrm{Aut}(\hat{G})$. Our nearest aim is to prove for such u and v that

$$\mathrm{V}(\hat{G},u,\hat{G},v) \le (n+4)/2, \tag{10}$$

where n is the order of $\hat{G}$.

By [13, Fact 1.6.1], $\mathrm{V}(\hat{G},u,\hat{G},v)$ is equal to the smallest l such that, for some r, in the game $\mathrm{E}{\scriptstyle\mathrm{HR}}_r^l(\hat{G},\hat{G})$ on vertex-disjoint copies of $\hat{G}$ Spoiler has a winning strategy from the position with u and v matched by a pair of identical pebbles. It therefore suffices to show that $l \le (n+4)/2$ pebbles are enough for Spoiler to win. We assume that u and v are equally colored for otherwise $l=1$ is enough. Take two colors that do not occur in $\hat{G}$, say, red and blue. Given $w \in V(\hat{G})$, define $\hat{G}_w$ to be the graph $\hat{G}[V(G)\setminus\{w\}]$ with the following additional coloring: the vertices that were adjacent to w are colored red and the remaining vertices are colored blue. Since mapping u to v cannot be extended to an automorphism of $\hat{G}$, we have $\hat{G}_u \not\cong \hat{G}_v$. It follows that Spoiler can win by keeping one pair of pebbles on u and v and playing $\mathrm{E}{\scriptstyle\mathrm{HR}}_{l-1}^{l-1}(\hat{G}_u,\hat{G}_v)$ with $l-1 = \mathrm{D}(n-1)$. Using the generalization of Theorem 1 to colored graphs, we conclude that Spoiler is able to win with $l \le 1+((n-1)+3)/2 = (n+4)/2$ pebbles thereby proving (10).

Let $\hat{\mathcal{S}}$ denote an analog of class $\mathcal{S}$ for colored graphs, that is, $\hat{\mathcal{S}}$ consists of all colored graphs $\hat{G}$ with $\sigma(\hat{G}) > (|\hat{G}|+3)/2$. We are now prepared to prove the bound (7). Assume that a graph G is not in $\mathcal{S}$ and $|G| \le n$. Consider an arbitrary coloration $\hat{G}$ of G. Taking into account (9) and (10), it is enough to estimate $\mathrm{V}(\hat{G})$. Obviously, $\sigma(\hat{G}) \le \sigma(G)$. It follows that $\hat{G} \notin \hat{\mathcal{S}}$ and, by Theorem 7, $\mathrm{V}(\hat{G}) \le (n+5)/2$ as required.

We now prove (8). Assume that a graph G is in $\mathcal{S}$ and $|G| = n$. Again, consider an arbitrary coloration $\hat{G}$ of G. If $\hat{G} \notin \hat{\mathcal{S}}$, then $\mathrm{V}(\hat{G}) \le (n+5)/2$ as before. If $\hat{G} \in \hat{\mathcal{S}}$, then, by Theorem 7, $\mathrm{V}(\hat{G}) = \sigma(\hat{G})+1 \le \sigma(G)+1$. By the definition of $\mathcal{S}$, $\sigma(G)+1 > (n+5)/2$ and we hence have the bound $V(\hat{G}) \le \sigma(G)+1$ for all $\hat{G}$. Taking into account also (10), we obtain $\hat{\mathrm{V}}(G) \le \sigma(G) + 1$. Since this bound is attained by $\mathrm{V}(\hat{G})$ for $\hat{G} = G$, we conclude that $\hat{\mathrm{V}}(G) = \sigma(G) + 1$ and $\hat{\mathrm{V}}(G) > (n + 5)/2$. $\square$

5 Computational complexity of Ehrenfeucht games

Pezzoli in [15] addresses the computational problem of determining, given two structures A, B and a number r, who of the players, Spoiler or Duplicator, has a winning strategy in the game $\mathrm{EHR}_r(A, B)$. She proves that, for structures over any fixed vocabulary containing at least one binary and one ternary relation, the problem is PSPACE-complete. For structures with a single binary relation, the complexity of this problem remains unknown. One can only say that it is at least as hard as recognition if two structures of the same order r are isomorphic. In the case of graphs, our results imply an extension of the latter observation.

Corollary 3. *The problem of determining, given two graphs G, G' and a number $r \ge (|G| + 5)/2$, who of the players has a winning strategy in the game $\mathrm{EHR}_r(G, G')$ is computationally equivalent to recognition if G and G' are isomorphic.*

Proof. The aforementioned reduction of the graph isomorphism recognition to the winner recognition is obvious: G and G' of the same order n are isomorphic iff Duplicator wins $\mathrm{EHR}_n(G, G')$.

The reduction in the other direction proceeds as follows. First decide if $G \cong G'$. If so, the winner is Duplicator. If not, decide if G is in $\mathcal{C}$. In the case that $G \notin \mathcal{C}$, the winner is Spoiler because $r \ge \mathrm{D}(G)$ by the bound for $\mathrm{D}(G)$ in Theorem 2 and the bound imposed on r. In the case that $G \in \mathcal{C}$, decide if the pair G, G' satisfies the condition (5), possibly with G and G' interchanged (here we again need the ability to test graph isomorphism). If (5) is false, the winner is Spoiler because $r \ge \mathrm{D}(G, G')$ by the bound for $\mathrm{D}(G, G')$ in Lemma 4 and the bound imposed on r. If (5) is true, compute $\mathrm{D}(G, G')$ by using Lemma 5. The winner is Duplicator if $r < \mathrm{D}(G, G')$ and Spoiler otherwise. $\square$

6 Conclusion

In this paper we have shown almost optimal upper bounds for first order logic definability over graphs. In our current work we are extending our techniques to other clases of interest, in particular to directed graphs and relational structures.

Acknowledgments

The third author is grateful to Georg Gottlob and members of his department at the Vienna University of Technology for their hospitality during this author's visit in the summer 2002.

References

[1] N. Alon, Tools from higher algebra. Chapter 32 of the *Handbook of Combinatorics*, 1749–1783. R. L. Graham, M. Grötschel, L. Lovász Eds. Elsevier Publ. (1995).

[2] N. Alon, P. Seymour, R. Thomas, A separator theorem for nonplanar graphs. *J. Am. Math. Soc.* 3(4):801–808 (1990).

[3] L. Babai, Automorphism groups, isomorphism, reconstruction. Chapter 27 of the *Handbook of Combinatorics*, 1447–1540. R. L. Graham, M. Grötschel, L. Lovász Eds. Elsevier Publ. (1995).

[4] L. Babai, On the complexity of canonical labelling of strongly regular graphs. *SIAM J. Comput.* 9:212–216 (1980).

[5] L. Babai, P. Erdős, S. M. Selkow, Random graph isomorphism. *SIAM J. Comput.* 9:628–635 (1980).

[6] L. Babai, L. Kučera, Canonical labelling of graphs in linear average time. *Proc. of 20th Ann. Symp. on Foundations of Computer Science* 39–46 (1979).

[7] J.-Y. Cai, M. Fürer, N. Immerman, An optimal lower bound on the number of variables for graph identification. *Combinatorica* 12(4):389–410 (1992).

[8] H.-D. Ebbinghaus, J. Flum. *Finite Model Theory.* Springer Verlag, 2nd rev. ed. (1999).

[9] M. Grohe, Fixed-point logics on planar graphs. In: *Proc. of the Ann. Conf. on Logic in Computer Science* 6–15 (1998).

[10] M. Grohe, Isomorphism testing for embeddable graphs through definability. In: *Proc. of the 32nd ACM Ann. Symp. on Theory of Computing (STOC)* 63–72 (2000).

[11] M. Grohe, J. Marino, Definability and descriptive complexity on databases of bounded tree-width. In: *Proc. of the 7t h International Conference on Database Theory*, Lecture Notes in Computer Science 1540, Springer-Verlag, 70–82 (1999).

[12] N. Immerman, *Descriptive complexity.* Springer-Verlag (1999).

[13] N. Immerman, E. Lander, Describing graphs: a first-order approach to graph canonization. In: *Complexity Theory Retrospective*, A. Selman Ed., Springer-Verlag, 59–81 (1990).

[14] L. Kučera, Canonical labelling of regular graphs in linear average time. *Proc. of 28th Ann. Symp. on Foundations of Computer Science* 271–279 (1987).

[15] E. Pezzoli, Computational complexity of Ehrenfeucht-Fraïssé games on finite structures. In: *Proc. of the CSL'98 Conf.*, G. Gottlob, K. Seyr Eds. Lecture Notes in Computer Science 1584, Springer-Verlag, 159–170 (1999).

[16] E. Pezzoli, *On the computational complexity of type-two functionals and logical games on finite structures.* Ph. D. thesis. Stanford University.

[17] J. Spencer. *The strange logic of random graphs.* Springer Verlag (2001).