Combinatorics, Probability and Computing (2007) 16, 375–400. © 2007 Cambridge University Press
doi:10.1017/S0963548306008376 First published online 23 January 2007 Printed in the United Kingdom

First-Order Definability of Trees and Sparse Random Graphs

TOM BOHMAN[1†], ALAN FRIEZE[1‡], TOMASZ ŁUCZAK[2§], OLEG PIKHURKO[1¶], CLIFFORD SMYTH[3], JOEL SPENCER[4] and OLEG VERBITSKY[5‖]

[1]Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213, USA
(e-mail: tbohman@moser.math.cmu.edu,
alan@random.math.cmu.edu, pikhurko@cmu.edu)

[2]Department of Discrete Mathematics, Adam Mickiewicz University, Poznań 61-614, Poland
(e-mail: tomasz@amu.edu.pl)

[3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
(e-mail: csmyth@math.mit.edu)

[4]Courant Institute, New York University, New York, NY 10012, USA
(e-mail: spencer@cims.nyu.edu)

[5]Institut für Informatik, Humboldt Universität Berlin, D-10099 Berlin, Germany
(e-mail: verbitsk@informatik.hu-berlin.de)

Received 30 October 2004; revised 28 September 2006

Let $D(G)$ be the smallest quantifier depth of a first-order formula which is true for a graph G but false for any other non-isomorphic graph. This can be viewed as a measure for the descriptive complexity of G in first-order logic.

We show that almost surely $D(G) = \Theta(\frac{\ln n}{\ln \ln n})$, where G is a random tree of order n or the giant component of a random graph $\mathcal{G}(n, \frac{c}{n})$ with constant $c > 1$. These results rely on computing the maximum of $D(T)$ for a tree T of order n and maximum degree l, so we study this problem as well.

1. Introduction

This paper deals with graph properties expressible in first-order logic. The vocabulary consists of variables, connectives ($\vee$, $\wedge$ and $\neg$), quantifiers ($\exists$ and $\forall$), and two binary

[†] Partially supported by NSF grant DMS-0401147.
[‡] Partially supported by NSF grant CCR-0200945.
[§] Partially supported by KBN grant 1 P03A 025 27.
[¶] Partially supported by the Berkman Faculty Development Fund, CMU, and NSF grant DMS-0457512.
[‖] Supported by an Alexander von Humboldt fellowship.

relations: the equality and the graph adjacency ($=$ and $\sim$ respectively). The variables denote vertices only, so we are not allowed to quantify over sets or relations. The notation $G \models A$ means that a graph G is a model of a *sentence* A (a first-order formula without free variables); in other words, A is true for the graph G. All sentences and graphs are assumed to be finite. The reader is referred to Spencer's book [12] (or to Kim, Pikhurko, Spencer and Verbitsky [5]) for more details.

A first-order sentence A *distinguishes* G from H if $G \models A$ but $H \not\models A$. Further, we say that A *defines* G if A distinguishes G from any non-isomorphic graph H. In other words, G is the unique (up to an isomorphism) finite model of A.

The *quantifier depth* $D(A)$ is the largest number of nested quantifiers in A. This parameter is closely related to the complexity of checking whether $G \models A$.

The main parameter we study is $D(G)$, the smallest quantifier depth of a first-order formula defining G. We call this graph invariant the *logical depth* of G. It was first systematically studied by Pikhurko, Veith and Verbitsky [10] (see also [11]). In a sense, a defining formula A can be viewed as the canonical form for G (except that A is not unique): in order to check whether $G \cong H$ it suffices to check whether $H \models A$. Unfortunately, this approach does not seem to lead to better isomorphism algorithms but this notion, being on the borderline of combinatorics, logic and computer science, is interesting on its own.

Within a short time-span various results on the values of $D(G)$ for order-n graphs appeared. The initial papers [10, 11] studied the maximum of $D(G)$ (the 'worst' case). The 'best' case is considered by Pikhurko, Spencer and Verbitsky [9, 8], while Kim, Pikhurko, Spencer and Verbitsky [5] obtained various results for random graphs.

Here we study these questions for trees and sparse random structures. Namely, the three main questions we consider are as follows.

Section 3: What is $D^{\mathrm{tree}}(n,l)$, the maximum of $D(T)$ over all trees of order at most n and maximum degree at most l?

Section 4: What is $D(G)$, where G is the giant component of a random graph $\mathcal{G}(n,\frac{c}{n})$ for constant $c > 1$?

Section 5: What is $D(T)$ for a random tree T of order n?

In all cases we determine the order of magnitude of the studied function. Namely, we prove that $D^{\mathrm{tree}}(n,l) = \Theta(\frac{l \ln n}{\ln l})$, and w.h.p. we have $D(G) = \Theta(\frac{\ln n}{\ln \ln n})$, whenever G is a random tree of order n or the giant component of a random graph $\mathcal{G}(n,\frac{c}{n})$ with constant $c > 1$. (The abbreviation *w.h.p.* stands for 'with high probability', *i.e.*, with probability $1 - o(1)$ as $n \to \infty$.) Moreover, for some cases involving trees we estimate the smallest quantifier depth of a defining formula up to a factor of $1 + o(1)$. For instance, we show that for a random tree T of order n we have w.h.p. $D(T) = (1 + o(1))\frac{\ln n}{\ln \ln n}$.

2. Further notation and terminology

Our main tool in the study of $D(G)$ is the *Ehrenfeucht game*. Its description can be found in Spencer's book [12] whose terminology we follow (or see [5, Section 2]), so here we sketch only basic ideas and definitions related to this concept.

Given two graphs G and G', the *Ehrenfeucht game* $\textsc{Ehr}_k(G, G')$ is a perfect information game played by two players, called *Spoiler* and *Duplicator*, and consists of k rounds,

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

where k is known in advance to both players. For brevity, let us refer to Spoiler as 'him' and to Duplicator as 'her'. In the ith round, $i = 1,\ldots,k$, Spoiler selects one of the graphs G and G' and marks one of its vertices by i; Duplicator must put the same label i on a vertex in the other graph. At the end of the game let $x_1,\ldots,x_k$ be the vertices of G marked $1,\ldots,k$ respectively, regardless of who put the label there; let $x'_1,\ldots,x'_k$ be the corresponding vertices in G'. Duplicator wins if the correspondence $x_i \leftrightarrow x'_i$ is a partial isomorphism, that is, we require that $\{x_i, x_j\} \in E(G)$ if and only if $\{x'_i, x'_j\} \in E(G')$ as well as that $x_i = x_j$ if and only if $x'_i = x'_j$. Otherwise, Spoiler wins.

The key relation is that $D(G, G')$, the smallest quantifier depth of a first-order sentence A distinguishing G from G', is equal to the smallest k such that Spoiler can win $\textsc{Ehr}_k(G, G')$. Also,

$$D(G) = \max_{G' \not\cong G} D(G, G') \tag{2.1}$$

(see, *e.g.*, [5, Lemma 1]).

The parameters $D(G)$ and $D(G, G')$, the Ehrenfeucht game, and the relationship between them generalize to arbitrary finite structures with finitely many relations, if the notion of a (partial) isomorphism is understood appropriately. We technically gain by considering *coloured graphs* which, in addition to the adjacency relation, have unary relations $U_1, U_2, \ldots$ and binary relations $B_1, B_2, \ldots$. If $U_i(x) = 1$, we say that a vertex x has colour i. It is supposed that $B_j(x, y) = 1$ only if x and y are adjacent. In this case we say that a directed edge (x, y) has colour j.

Note that the B_j's are not supposed to be symmetric, that is, for an edge $\{x, y\}$ we consider two of its orientations, (x, y) and (y, x), whose colours may be different. Although the set of colours is potentially infinite, a coloured graph is supposed to have only finitely many colours. When the Ehrenfeucht game is played on coloured graphs, Duplicator must additionally preserve the colours of the selected vertices and of all edges between them.

Colourings can be useful even if we prove results for uncoloured graphs. For example, if $x \in V(G)$ and $x' \in V(G')$ were selected in some round, then, without changing the outcome of the remaining game, we can remove x and x' from G and G' respectively, provided we colour their neighbours with a new colour. (Note that in an optimal strategy of Spoiler, there is no need to select the same vertex twice if $k \leqslant \max(v(G), v(G'))$.)

We will also use the following fact, which can be easily deduced from the general theory of the Ehrenfeucht game. Let $x, y \in V(G)$ be distinct vertices. Then the smallest quantifier depth of a first-order formula Φ with one free variable such that $G, x \models \Phi$, but $G, y \not\models \Phi$, is equal to the minimum k for which Spoiler can win the $(k+1)$-round game $\textsc{Ehr}_{k+1}(G, G)$, where the vertices $x_1 = x$ and $x'_1 = y$ were selected in the first round.

In this paper ln denotes the natural logarithm, while the logarithm base 2 is written as $\log_2$. We also assume everywhere that n is sufficiently large to satisfy all stated inequalities.

3. General trees

Let $D^{\mathrm{tree}}(n, l)$ be the maximum of $D(T)$ over all coloured trees of order at most n and maximum degree at most l. Note that the number of colours does not appear in this definition. The maximum exists because $D(G) \leqslant n + 1$ for every coloured graph G on at

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

most n vertices. For some ranges of n, l we are able to compute $D^{\text{tree}}(n, l)$ asymptotically. At the end of the section we discuss the remaining values of n, l.

Theorem 3.1. *Let both l and $\ln n / \ln l$ tend to infinity. Then*

$$D^{\text{tree}}(n, l) = \left(\frac{1}{2} + o(1)\right) \frac{l \ln n}{\ln l}. \tag{3.1}$$

In fact, the lower bound can be achieved by uncoloured trees.

Theorem 3.2. *Let an integer $t \geqslant 1$ be fixed. Suppose that $l, n \to \infty$ are such that $n \geqslant l^t$ but $n = o(l^{t+1})$. Then*

$$D^{\text{tree}}(n, l) = \left(\frac{t+1}{2} + o(1)\right) l.$$

In fact, the lower bound can be achieved by uncoloured trees.

In order to prove Theorems 3.1 and 3.2 we need some preliminary results. Let $\text{dist}_G(x, y)$ denote the distance in G between $x, y \in V(G)$.

Lemma 3.3. *Suppose that in the Ehrenfeucht game on coloured graphs G and G', some vertices $x, y \in V(G)$ and $x', y' \in V(G')$ were selected in the same rounds. Let I be a set of colours. Suppose that there is an xy-path P of length k with no colours from I, while any $x'y'$-path of length at most k, if such exists, has a vertex with a colour in I.*

Then Spoiler can win in at most $\lceil \log_2 k \rceil$ extra moves, playing all of the time inside G.

Proof. We prove the claim by induction on k. If $k = 1$, then Spoiler has already won, so assume that $k \geqslant 2$. Spoiler selects a vertex $w \in V(G)$ which is a *middle vertex* of P, that is, $k_1 = \text{dist}_P(x, w)$ and $k_2 = \text{dist}_P(y, w)$ differ at most by one. Suppose that Duplicator responds with $w' \in G'$. Assume that no colour from I appears at w' for otherwise Spoiler has already won. It is impossible that G' contains an $x'w'$-path of length at most k_1 and a $y'w'$-path of length at most k_2, with no vertex there having a colour from I.

If, for example, the latter does not exist, then we apply induction to $y, w \in G$ and $y', w' \in G'$. The required bound follows by observing that $k_1, k_2 \leqslant \lceil \frac{k}{2} \rceil$. □

Lemma 3.4. *Let T be a tree of order n and let T' be a graph which is not a tree. Then $D(T, T') \leqslant \log_2 n + 4$.*

Proof. If T' is not connected, Spoiler selects two vertices $x', y' \in T'$ from different components. Then he switches to T and applies Lemma 3.3, winning in at most $\log_2 n + 3$ moves in total.

Otherwise, let $C' \subset T'$ be a cycle of the shortest length k. If $k \geqslant 2n$, then Spoiler picks two vertices x', y' at distance at least n in C' (and hence in T'). But the diameter of T is at most $n - 1$. Spoiler switches to T and wins in at most $\log_2 n + 3$ moves by Lemma 3.3.

If $k < 2n$, then Spoiler selects some three vertices x', z', y' of C', every two at distance at most n. Assume that of Duplicator's replies in the tree T, z lies between x and y. Spoiler

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

applies Lemma 3.3 to $G = T' - z'$ and $G' = T - z$, making at most $3 + (\log_2 n + 1)$ moves in total. □

We are ready to prove Theorem 3.1. For future use it will be convenient to have an upper bound which is valid for *any* l. The following lemma serves this purpose (except perhaps some $l \leqslant 8$, when we can use the bound $D^{\mathrm{tree}}(n, l) \leqslant D^{\mathrm{tree}}(n, 9)$).

Lemma 3.5. *There is a constant C such that for any $9 \leqslant l \leqslant n - 1$ we have*

$$D^{\mathrm{tree}}(n, l) \leqslant \frac{l \ln n}{2 \ln(l/2)} + \frac{3l}{2} + 3 \log_2 n + C. \tag{3.2}$$

Proof. Let T be any tree of order at most n and maximum degree at most l. Let T' be an arbitrary coloured graph not isomorphic to T. By Lemma 3.4 we can assume that T' is a tree. We also assume that T and T' have the same diameter because otherwise Spoiler wins in less than $\log_2 n + 1$ moves by Lemma 3.3.

It is easy to show (see, *e.g.*, Ore [7]) that T contains a vertex $x \in T$ such that any component of $T - x$ has order at most $\frac{n}{2}$. We call such a vertex a *median* of T. Spoiler selects this vertex x; let Duplicator reply with x'. We can assume that the degrees of x and x' are the same: otherwise Spoiler can exhibit this discrepancy in at most $l + 1$ extra moves.

We update the colouring of each component of $T - x$ and $T' - x'$ as follows. Let C be a component of $T - x$ and y be the neighbour of x in C (with the components of $T' - x'$ we proceed similarly). In addition to the colours that already exist in C, the vertex y receives a new colour which consists of three components: the set of j such that $B_j(x, y) = 1$, the set of j such that $B_j(y, x) = 1$, and the number of the round in which x and x' are selected.

As $T \not\cong T'$, some component C_1 has different multiplicities m_1 and m_1' in $T - x$ and $T' - x'$. As $d(x) = d(x')$, we have at least two such components. Assume that for C_1 and C_2 we have $m_1 > m_1'$ and $m_2 < m_2'$. By the condition on the maximum degree, $m_1' + m_2 \leqslant l - 1$. Hence, $\min(m_1', m_2) \leqslant \frac{l-1}{2}$. Let us assume, for example, that $m_1' \leqslant \frac{l-1}{2}$. Spoiler chooses the neighbours of x inside any $m_1' + 1$ C_1-components of $T - x$. It must be the case that some vertices $y \in V(T)$ and $y' \in V(T')$ have been selected in the same round, so that y lies in a C_1-component $F \subset T - x$ while y' lies in a component $F' \subset T' - x$ not isomorphic to C_1. Let n_1 be the number of vertices in F. By the choice of x, $n_1 \leqslant \frac{n}{2}$.

Now, Spoiler restricts his moves to $V(F) \cup V(F')$. If Duplicator moves outside this set, then Spoiler applies Lemma 3.3 to $T - x$ and $T' - x'$, winning in at most $\log_2 n + 1$ moves. Otherwise Spoiler uses the recursion applied to F.

Let $f(n, l)$ denote the largest number of moves (over all coloured trees T, T' with $v(T) \leqslant n$, $\Delta(T) \leqslant l$, and $T \not\cong T'$) that Duplicator can survive against the above strategy, with the additional restriction that a situation where Lemma 3.3 can be applied never occurs and always $d(x) = d(x')$. Clearly,

$$D^{\mathrm{tree}}(n, l) \leqslant f(n, l) + \log_2 n + l + O(1). \tag{3.3}$$

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

As $m_1 \leqslant \frac{n-1}{n_1}$, we get the following recursive bound on f:

$$f(n,l) \leqslant \max\left\{2 + \min\left(\tfrac{l-1}{2}, \tfrac{n-1}{n_1}\right) + f(n_1,l) : 1 \leqslant n_1 \leqslant \tfrac{n}{2}\right\}. \tag{3.4}$$

Denoting $n_0 = n$ and unfolding (3.4) as long as $n_i \geqslant 1$, say s times, we obtain that $f(n,l)$ is bounded by the maximum of

$$2s + \sum_{i=1}^{s} \min\left(\frac{l-1}{2}, \frac{n_{i-1}}{n_i}\right), \tag{3.5}$$

over all sequences $n_1,\ldots,n_s$ such that

$$1 \leqslant n_i \leqslant \frac{n_{i-1}}{2}, \quad i \in [s]. \tag{3.6}$$

Note that the restrictions (3.6) force s to be at most $\log_2 n$. Let us maximize (3.5) over all $s \in \mathbb{N}$ and real n_i's satisfying (3.6).

It is easy to see that for the optimal sequence we have $2 \leqslant \frac{n_{i-1}}{n_i} \leqslant \frac{l-1}{2}$, $i \in [s]$. Moreover, both these inequalities can be simultaneously strict for at most one index i. Indeed, suppose on the contrary that for two indices $0 \leqslant i < j < s$ we have $2 < n_i/n_{i+1} < \frac{l-1}{2}$ and $2 < n_j/n_{j+1} < \frac{l-1}{2}$. Define a new sequence: $n'_h = n_h$ if $h \leqslant i$ or $h > j$, while $n'_h = \gamma n_h$ for $i < h \leqslant j$. If $\gamma = 1$, then we obtain the same sequence. Note that $\frac{n'_h}{n'_{h+1}} = \frac{n_h}{n_{h+1}}$ for any h except $h = i$ or $h = j$. So, we can slightly perturb γ either way, without violating (3.6). The right-hand side of (3.5), as a function of γ in a small neighbourhood of $\gamma = 1$, is of the form $a\gamma + \frac{b}{\gamma} + c$ with $a, b > 0$. But this function is strictly convex, so it cannot attain its maximum at $\gamma = 1$, a contradiction.

Let t be the number of times we have $n_{i-1} = 2n_i$. The bound (3.5) reads

$$f(n,l) - 2\log_2 n \leqslant 2t + (s-t)\,\frac{l-1}{2}. \tag{3.7}$$

Given that $2^t(\frac{l-1}{2})^{s-t-1} \leqslant n$, the right-hand side of (3.7) is maximized over reals $t \geqslant 0$ and $0 \leqslant s \leqslant \log_2 n$ if $t = 0$ and $(\frac{l-1}{2})^{s-1} = n$, when the extremal value is

$$\left(\frac{\ln n}{\ln((l-1)/2)} + 1\right)\frac{l-1}{2} \leqslant \frac{l \ln n}{2\ln(l/2)} + \frac{l}{2}.$$

This implies the upper bound (3.2) by (3.3) and (3.7). □

Proof of Theorem 3.1. The upper bound follows from Lemma 3.5. It remains to prove the lower bound. Let $k = \lfloor l/2 \rfloor$. Define $G_0 = K_{1,l-1}$ and $G'_0 = K_{1,l-2}$. Let the roots $r_0 \in V(G_0)$, $r'_0 \in V(G'_0)$ be the vertices of high degree. Define inductively on i the following graphs. G_i is obtained by taking k copies of G_{i-1} and $k-1$ copies of G'_{i-1}, pairwise vertex-disjoint, plus the root r_i connected to the root of each copy of G_{i-1} and G'_{i-1}. We have $d(r_i) \leqslant l-1$. The graph G'_i is defined in a similar way except that we take $k-1$ copies of G_{i-1} and k copies of G'_{i-1}. Let i be the largest index such that $\max(v(G_i), v(G'_i)) \leqslant n$.

Let us disregard all roots, *i.e.*, view G_j and G'_j as usual (uncoloured) graphs. Note that the trees G_i and G'_i are non-isomorphic because for every j we can identify the level-j roots as the vertices at distance $j+1$ from some leaf.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Define $g_j = (k-1)j + l - 2$, $j \in [0, i]$. Let us show by induction on j that Duplicator can survive at least g_j rounds in the Ehrenfeucht game on (G_j, G'_j). This is clearly true for $j = 0$. Let $j \geqslant 1$. If Spoiler claims one of r_j, r'_j then Duplicator selects the other. If Spoiler selects a vertex in a graph from the 'previous' level, for example $F \subset G_j$ with $F \cong G'_{j-1}$, then Duplicator chooses an $F' \subset G'_i$, $F' \cong G'_{j-1}$ and keeps the isomorphism between F and F'. So any further moves of Spoiler inside $V(F) \cup V(F')$ will be useless and we can ignore F and F'. Thus it takes Spoiler at least $k-1$ moves before we are down to the pair (G_{j-1}, G'_{j-1}), which proves the claim.

Thus we have $D(G_i) \geqslant D(G_i, G'_i) \geqslant g_i = (\frac{1}{2} + o(1)) \frac{l \ln n}{\ln l}$, finishing the proof. If we wanted to construct a tree with *exactly* n vertices, then we could just affix a path of length $n - v(G_i)$ to the roots of both G_i and G'_i; the obtained graphs, H_i and H'_i, would satisfy $v(H_i) = n$ and $D(H_i) \geqslant D(H_i, H'_i) \geqslant g_i$. □

Remark. Verbitsky [14] proposed a different argument to estimate $D^{\mathrm{tree}}(n, l)$, which gives weaker bounds than those in Theorem 3.1 but can be applied to other classes of graphs with small separators.

Proof of Theorem 3.2. Since the arguments are similar to those of Theorem 3.1 we will be brief.

Although the stated upper bound is slightly stronger than the one given by Lemma 3.5, Spoiler uses the same strategy as before. Namely, he chooses a median $x \in T$ and of two possible multiplicities, summing up to at most l, chooses the smaller. Let $m_1 + 1, m_2 + 1, \ldots, m_k + 1$ be the number of moves for each selected median. Since $m_i < n_{i-1}/n_i$ for the n_i's as in the proof of Lemma 3.5, we have $n \geqslant \prod_{i=1}^{k} m_i$. Also, we have $k \leqslant \log_2 n$ because we always choose a median. Given these restrictions, the inequalities $m_i \leqslant (l+1)/2$, $i \in [k-1]$, and $m_k \leqslant l$, the sum $\sum_{i=1}^{k} m_i$ is maximized if $m_k = l$ and as many as possible m_j with $j \in [m-1]$ are maximum possible (*i.e.*, $m_j = (l+1)/2$). We thus factor out $(l+1)/2$ at most $t-1$ times until the remaining terms have the product (and so the sum) $o(l)$. Thus,

$$\sum_{i=1}^{k} (m_i + 1) \leqslant \sum_{i=1}^{k} m_i + \log_2 n \leqslant l + \frac{(t-1)l}{2} + o(l),$$

as required.

The lower bound is given by the construction in the proof of Theorem 3.1: we have $g_{t-1} = (\frac{t+1}{2} + o(1))l$ and, by induction on j, the tree G_j has at most l^j vertices. □

Theorems 3.1 and 3.2 do not cover all the possibilities for n, l. The asymptotic computation in the remaining cases seems rather messy. However, the order of magnitude of $D^{\mathrm{tree}}(n, l)$ is easy to compute with what we already have. For example, Theorem 3.2 implies that for $n = \Theta(l^t)$ with fixed integer $t \geqslant 1$ we have $D^{\mathrm{tree}}(n, l) = \Theta(l)$. If $l \geqslant 2$ is constant, then $D^{\mathrm{tree}}(n, l) = \Theta(\ln n)$, where the lower bound follows by noting that for the n-path we have $D(P_n) = \log_2 n + O(1)$ (see [12, Section 2.1]) and the upper bound follows from Lemma 3.5.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

4. The giant component

Let us recall that if $p = c/n$ with constant $c > 1$, then w.h.p. in a random graph $H \in \mathcal{G}(n,p)$ there exists a unique component containing a positive fraction of all vertices. The complexity of such a *giant component* is given by the following result.

Theorem 4.1. *Let $c > 1$ be a constant, $p = c/n$, and let G be the giant component of a random graph $H \in \mathcal{G}(n,p)$. Then w.h.p.*

$$D(G) = \Theta\left(\frac{\ln n}{\ln \ln n}\right). \tag{4.1}$$

This result has the following consequence.

Theorem 4.2. *Let $c > 0$ be a constant, $p = \frac{c}{n}$, and $H \in \mathcal{G}(n,p)$. Then w.h.p.*

$$D(H) = (\mathrm{e}^{-c} + o(1))\, n. \tag{4.2}$$

Proof. The proof is an easy modification of the arguments in [5, Theorem 19], where the validity of (4.2) was established for $c \leqslant 1.19\ldots$.

The lower bound in (4.2) comes from considering the graph H' obtained from H by adding an isolated vertex (and noting that w.h.p. H has $(\mathrm{e}^{-c} + o(1))\, n$ isolated vertices).

On the other hand, let H' be any graph non-isomorphic to H. We can assume that Duplicator preserves the connectivity relation for otherwise Spoiler wins in extra $\log_2 n + O(1) = o(n)$ moves.

For a connected graph F, let $c_F(H)$ be the number of connectivity components in H isomorphic to F. Since $H \ncong H'$, there is an F with $c_F(H) \neq c_F(H')$, say $c_F(H) < c_F(H')$. Spoiler selects some $c_F(H) + 1$ F-components of H' and picks one vertex in each of them. By our assumptions, in some round the players must select $x' \in H'$ and $x \in H$ belonging to non-isomorphic components C' and C (namely, $C' \cong F \ncong C$). Now Spoiler can win the game in at most $D(C, C')$ extra moves (assuming that Duplicator preserves connectivity). But $D(C, C') \leqslant D(C)$, which is at most $O(\frac{\ln n}{\ln \ln n})$ if C is the giant component by Theorem 4.1 and at most $v(C) + 1 = o(n)$ otherwise.

It was shown in [5, Theorem 19] by using a theorem of Barbour [1] (Theorem 5.6 in Bollobás [2]) that w.h.p. $c_F(H) + v(F) \leqslant (\mathrm{e}^{-c} + o(1))n$ for any F except perhaps the giant component. This finishes the proof of Theorem 4.2. □

4.1. Upper bound

The structure of the giant component is often characterized using its core and kernel (*e.g.*, see Janson, Łuczak and Ruciński [4, Section 5]). We follow this approach in the proof of the upper bound in (4.1). Thus, we first bound $D(G)$ from above for a graph G with small diameter whose kernel fulfils some 'sparseness' conditions. Then, we show that these conditions hold w.h.p. for the kernel of the giant component of a random graph.

4.1.1. Bounding $D(G)$ using the kernel of G. The *core* C of a graph G is obtained by removing, consecutively and as long as possible, vertices of degree at most 1. If G is not a forest, then C is non-empty and $\delta(C) \geqslant 2$.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

First we need an auxiliary lemma which is easily proved, in a similar way to the auxiliary lemmas in Section 3, by the path-halving argument.

Lemma 4.3. *Let G, G' be graphs. Suppose that $x \in V(G)$ and $x' \in V(G')$ have been selected in the same round so that G contains some cycle P of length at most k with $x \in V(P)$ while G' does not contain such a short cycle through x'. Then Spoiler can win in at most $\log_2 k + O(1)$ moves, playing all the time inside G.*

Proof. Spoiler chooses y, z, the neighbours of x in the cycle P. The distance between y and z in $G - x$ is at most $k - 2$, which cannot be true for Duplicator's replies y' and z' in $G' - x'$. Now, Spoiler applies the strategy of Lemma 3.3. □

Lemma 4.4. *Let G, G' be graphs and C, C' be their cores. If Duplicator does not respect the core, then Spoiler can win in at most $\log_2 d + O(1)$ extra moves, where d is the diameter of G.*

Proof. Assume that $\mathrm{diam}(G') = \mathrm{diam}(G)$ for otherwise Spoiler (unconditionally) wins in at most $\log_2 d + O(1)$ moves by Lemma 3.3. Suppose that, for example, some vertices $x \in V(C)$ and $x' \in V(G') \setminus V(C')$ have been selected.

If x lies on a cycle $C_1 \subset C$, then we can find such a cycle of length at most $2d + 1$. Of course, G' cannot have a cycle containing x', so Spoiler wins by Lemma 4.3 in $\log_2(2d + 1) + O(1)$ moves, as required.

Suppose that x does not belong to a cycle. Then G contains two vertex-disjoint cycles C_1, C_2 connected by a path P containing x. Choose such a configuration which minimizes the length of $P \ni x$. Then the length of P is at most $2d$ (in fact, at most d). Spoiler selects the branching vertices $y_1 \in V(C_1) \cap V(P)$ and $y_2 \in V(C_2) \cap V(P)$. If some Duplicator's reply y_i' is not on a cycle, we are done again by Lemma 4.3. So assume there are cycles $C_i' \ni y_i'$. In G we have

$$\mathrm{dist}(y_1, y_2) = \mathrm{dist}(y_1, x) + \mathrm{dist}(y_2, x). \tag{4.3}$$

As $x' \notin C'$, any shortest $x'y_1'$-path and $x'y_2'$-path enter x' via the same edge $\{x', z'\}$. But then

$$\mathrm{dist}(y_1', y_2') \leqslant \mathrm{dist}(y_1', z') + \mathrm{dist}(y_2', z') = \mathrm{dist}(y_1', x') + \mathrm{dist}(y_2', x') - 2. \tag{4.4}$$

By (4.3) and (4.4), the distances between x, y_1, y_2 cannot all be equal to the distances between x', y_1', y_2'. Spoiler can demonstrate this in at most $\log_2(\mathrm{dist}(y_1, y_2)) + O(1)$ moves, as required. □

In order to state our upper bound on $D(G)$ we have to define a number of parameters of G. In outline, we try to show that any distinct $x, y \in V(C)$ can be distinguished by Spoiler reasonably fast. This would mean that each vertex of C can be identified by a first-order formula of small quantifier depth. Note that G can be decomposed into the core and a number of trees T_x, $x \in V(C)$, rooted at vertices of C. Thus, by specifying which pairs of vertices of C are connected and describing each T_x, $x \in V(C)$, we completely define

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

G. However, we have one unpleasant difficulty that not all pairs of points of C can be distinguished from one another. For example, we may have a pendant triangle on $\{x, y, z\}$ with $d(x) = d(y) = 2$, in which case the vertices x and y are indistinguishable. However, we show that w.h.p. we can distinguish any two vertices of degree 3 or more in C, which suffices for our purposes.

Let us give all the details. For $x \in V(C)$, let $T_x \subset G$ denote the tree rooted at x, *i.e.*, T_x is the component containing x in the forest obtained from G by removing all edges of C. Let

$$t = \max\{D(T_x) : x \in V(C)\}, \tag{4.5}$$

where $D(T_x)$ is taken with respect to the class of graphs with one root (*i.e.*, a vertex of a special colour).

Let the *kernel* K of G be obtained from C by the *serial reduction* where we repeat as long as possible the following step: if C contains a vertex x of degree 2, then remove x from $V(C)$ but add the edge $\{y, z\}$ to $E(C)$ where y, z are the two neighbours of x. Note that K is a *multigraph* (that is, it may contain loops and multiple edges). We agree that each loop contributes 2 to the degree; thus the minimal degree $\delta(K) \geqslant 3$.

We view the core C as a vertex-coloured graph with the colour $c(x)$ of a vertex $x \in V(C)$ being the isomorphism type of the rooted tree T_x. Also, we colour the oriented edges and loops of K as follows. An oriented edge or loop $\overrightarrow{e}$ of K corresponds to the directed path $P_{\overrightarrow{e}}$ in C, say connecting x to y with $x = y$ if e is a loop. The colour of $\overrightarrow{e}$ is the sequence $c(P_{\overrightarrow{e}}) = (c(x), \ldots, c(y))$ of colours that we see in the core as we go along this path. Note that we do not colour the vertices of K. This edge-colouring of K is redundant but it has the desired property that the kernels (if non-empty) of any two non-isomorphic graphs are non-isomorphic as coloured multigraphs.

We have not yet defined the Ehrenfeucht game, *etc*., for multigraphs. On the intuitive level, the corresponding notions are fairly obvious. Here is the formal description for rigour's sake. We regard the multigraph K as a vertex- and edge-coloured *simple graph* as follows. The colour of (x, y) with $\{x, y\} \in E(K)$ is the multiset (that is, multiplicities are noted) consisting of $c(P_{\overrightarrow{e}})$ over oriented edges $\overrightarrow{e}$ connecting x to y in K. The colour of a vertex $x \in V(K)$ is the multiset of $c(P_{\overrightarrow{e}})$ over all loops e at x and two possible orientations $\overrightarrow{e}$ of each loop. Again, non-isomorphic edge-coloured multigraphs produce non-isomorphic coloured graphs and the usual graph concepts of definability, Ehrenfeucht game, *etc*., apply.

Let

$$u = \Delta(G) \quad \text{and} \quad d = \operatorname{diam}(G). \tag{4.6}$$

It follows that each edge of K corresponds to the path P in C of length at most $2d$. Assume that $u \geqslant 9$.

We now introduce an integer parameter h assuming that K satisfies certain conditions.

Assumption 1. Every set of $v \leqslant 6h + 5$ vertices of K spans at most v edges in K. (Roughly speaking, we do not have two short cycles close together.)

For $\{x, y\} \in E(K)$ let $A_{x,y}$ be the set of vertices obtained by doing breadth-first search in $K - x$ starting with y until the process dies out or, after we have added a whole level,

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

we reach at least 2^h vertices in total. Let $K_{x,y} = K[A_{x,y} \cup \{x\}]$ be a graph with two special roots x and y.

The *height* of $z \in V(K_{x,y})$ is the distance in $K - x$ between z and y. It is easy to deduce from the condition on short cycles and the inequality $\delta(G) \geqslant 3$ that each $K_{x,y}$ has at most one cycle (including loops) and the maximum height is at most h. In fact, the process dies out only in the case if y is an isolated loop in $K - x$. For $\{x, y\} \in E(K)$ let $G_{x,y}$ be the subgraph of G corresponding to $K_{x,y}$. We view $K_{x,y}$ and $G_{x,y}$ as having two special *roots* x and y, each root having its unique colour.

Here we impose another condition on G and h.

Assumption 2. Let $\{x, x'\}$, $\{y, y'\} \in E(K)$. If $K_{x,x'}$ and $K_{y,y'}$ both have order at least 2^h and $A_{x,x'} \cap A_{y,y'} = \emptyset$, then the rooted graphs $G_{x,x'}$ and $G_{y,y'}$ are not isomorphic.

Let us define

$$\begin{aligned} b_0 &= \frac{u \ln(u2^h)}{2\ln(u/2)} + \frac{3u}{2} + 3\log_2 u + 3h + \log_2 d, \\ b_1 &= b_0 + u + 2\log_2 d, \\ b &= b_1 + t + u + 2\log_2 d. \end{aligned}$$

Lemma 4.5. *Under Assumptions 1 and 2, we have* $D(G) \leqslant b + O(1)$.

Proof. Since the proof is rather complicated we have split it into a sequence of claims. As the result, the bound we prove, namely $D(G) \leqslant b + O(1)$, is slightly worse than the best bound given by this method. Since we were not able to obtain the asymptotic result in Theorem 4.1 anyway, we present the weaker bound for the sake of the clarity of exposition.

Let $G' \ncong G$. Let C', K' be its core and kernel. We can assume that $\Delta(G') = u$ and its diameter is d, for otherwise Spoiler easily wins in $u + 2$ or $\log_2 d + 3$ moves.

By Lemma 4.4 it is enough to show that Spoiler can win the Ehrenfeucht (G, G')-game in at most $b - \log_2 d + O(1)$ moves provided Duplicator always respects $V(C)$ and $V(K)$ (recall that the latter consists of the vertices in C having degree at least 3). Call this game $\mathscr{C}$.

As $G \ncong G'$, we have $K \ncong K'$ (as coloured graphs). Let $\mathscr{K}$ denote the Ehrenfeucht game on K and K'.

Claim 1. If Spoiler can win the game $\mathscr{K}$ in m moves, then he can win $\mathscr{C}$ in at most $m + t + u + \log_2 d + O(1)$ moves.

Proof of claim. We can assume that each edge of K' corresponds to a path in G' of length at most $2d + 1$: otherwise Spoiler selects a vertex of C' at distance at least $d + 1$ from any vertex of K' and wins in $\log_2 d + O(1)$ moves.

Spoiler plays according to his $\mathscr{K}$-strategy by making moves inside $V(K) \subset V(G)$ or $V(K') \subset V(G')$. By the definition of $\mathscr{C}$, Duplicator's replies are inside $V(K) \cup V(K')$, so they correspond to replies in the $\mathscr{K}$-game. In at most m moves, Spoiler wins the $\mathscr{K}$-game.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Of a few similar cases, assume that Spoiler achieved that the multisets of coloured edges between some selected vertices $x \neq y$ of $V(K)$ and $x' \neq y'$ of $V(K')$ are different.

In at most u moves, Spoiler can either win or select a vertex $z \in V(C) \setminus V(K)$, the neighbour of x in an xy-path P, such that the Duplicator's reply z' either is not in an $x'y'$-path or its path $P' \ni z'$ has a different colouring from P. In the former case, Spoiler wins by Lemma 3.3: in G the vertices y and z are connected by a path of length at most $2d$ that avoids any other vertex of K, while this does not hold for G', y', z'.

Consider the latter case. Assume that $|P| = |P'|$, for otherwise we are done by Lemma 3.3. Spoiler selects $w \in P$ such that, for the vertex $w' \in P'$ with $\mathrm{dist}_P(w, x) = \mathrm{dist}_{P'}(w', x')$, we have $T_w \ncong T'_{w'}$. If Duplicator does not reply with w', then she has violated distances in one way or another and Spoiler wins by Lemma 3.3. Otherwise Spoiler needs at most t extra moves to win the game $\mathscr{T}$ on $(T_w, T'_{w'})$ (and at most $\log_2 d + O(1)$ extra moves to catch Duplicator if she does not respect $\mathscr{T}$). ∎

It remains to bound $D(K)$, where K is the coloured graph, by $b_1 + O(1)$. This requires a few preliminary facts.

Claim 2. For any $\{x, x'\} \in K$ we have $D(K_{x,x'}) \leqslant b_0 + O(1)$ in the class of coloured graphs with two roots.

Proof of claim. Let $T = K_{x,x'}$ and $T' \ncong T$. If T is a tree, then we just apply Lemma 3.5 using the bound of $u2^h$ for the order and the bound of u for the maximum degree.

Otherwise, Spoiler first selects a vertex $z \in T$ which lies on the (unique) cycle. We have at most $u - 1$ components in $T - z$, viewing each as a coloured tree where one extra colour marks the neighbours of z. As $T \ncong T'$, in at most u moves we can restrict our game to one of the components. (If Duplicator does not respect components, she loses in at most $\log_2 d + O(1)$ extra moves.) Now, one of the graphs is a coloured tree, and Lemma 3.5 applies. ∎

Claim 3. For every two distinct vertices $x, y \in V(K)$, there is a first-order formula $\Phi_{x,y}$ in the language of coloured graphs, with one free variable and quantifier depth at most $b_1 + O(1)$, such that $K, x \models \Phi_{x,y}$ and $K, y \not\models \Phi_{x,y}$.

Proof of claim. To prove the existence of $\Phi_{x,y}$ we have to describe Spoiler's strategy, where he has to distinguish (K, x) and (K, y) for given distinct $x, y \in K$.

If the multiset of isomorphism classes $K_{x,x'}$, over $\{x, x'\} \in E(K)$ is not equal to the multiset $\{K_{y,y'} : \{y, y'\} \in E(K)\}$, then Spoiler wins in at most $b_0 + u + \log_2 h + O(1) \leqslant b_1 + O(1)$ moves by Claim 2. Indeed, Spoiler ensures that in at most $u + 1$ moves some vertices x' and y' are selected in the same round so that $K_{x,x'} \ncong K_{y,y'}$. Let $r \leqslant h$ be the height of $K_{x,x'}$ and let $\bar{K}_{y,y'}$ be the subgraph of $K_{y,y'}$ induced by the vertices of height at most r. Then $K_{x,x'} \ncong \bar{K}_{y,y'}$ and Spoiler can apply the strategy given by Claim 2. If Duplicator moves outside these graphs, then Spoiler wins in at most $\log_2 h + O(1)$ extra moves.

Assume now that the above multisets are equal. We show that in this case Spoiler wins even faster. By the inequality $\delta(K) \geqslant 3$ and Assumption 1, there are x' and y' such

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

that $K_{x,x'}$ and $K_{y,y'}$ have more than two vertices and are isomorphic. This implies an isomorphism $G_{x,x'} \cong G_{y,y'}$. By Assumption 2, the latter implies that $A_{x,x'} \cap A_{y,y'} \neq \emptyset$. As the height of any $K_{a,b}$ is at most h, we conclude that

$$\mathrm{dist}_K(x,y) \leqslant 2h+2. \tag{4.7}$$

It follows that neither x nor y is a loop. Indeed, otherwise both must be loops (or Spoiler has already won). But then Assumption 1 implies that $\mathrm{dist}_K(x,y) > 6h+4$, a contradiction to (4.7).

At most one neighbour of x can be a loop, for otherwise we get 3 vertices spanning 4 edges. The same holds for y. By (4.7) and Assumption 1, at least one of x, y has no loop among its neighbours. If this is so for exactly one of x, y, then Spoiler wins in one more move. So assume that no K-neighbour of x and y is a loop.

Let P be a shortest xy-path in K. Let x_1, x_2, x_3 and y_1, y_2, y_3 be three K-neighbours of x and y respectively such that $K_{x,x_i} \cong K_{y,y_i}$. We can additionally assume that both $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$ intersect P. Recall that

$$A_{x,x_i} \cap A_{y,y_i} \neq \emptyset, \quad i \in [3], \tag{4.8}$$

and the height of the corresponding subgraphs of K is at most h.

It is impossible to have three (not necessarily disjoint) paths each of length at most $2h+2$ of the form $(x, x_i, \ldots, y_i, y)$, $i \in [3]$, as this would give $v \leqslant 6h+5$ vertices spanning at least $v+1$ edges in K, a contradiction to Assumption 1. It follows that the length of P is at most h (so that one path may be used to ensure (4.8) for more than one index i). Still, P alone can take care of at most two such intersections so, up to a symmetry, one of the following two cases takes place.

Case 1. In K there is a cycle of length at most $h+(2h+2)$ containing the edges xx_1, xx_2, yy_1, yy_3.

Note that for $i = 1,2$ we have $A_{x,x_3} \cap A_{x,x_i} = \emptyset$ for otherwise we would have two short cycles close together. Thus $K_{x,x_3} \ncong K_{x,x_i}$.

Spoiler selects x_1 and x_2. If Duplicator does not reply with $\{y_1, y_3\}$, then Spoiler wins in extra $\log_2 h + O(1) \leqslant \log_2 d + O(1)$ moves, using the fact that x, x_1, x_2 lie on a cycle of length at most $3h+2$ (*cf.* the proof of Lemma 3.4). So, assume that one of Duplicator's replies is y_3. But $K_{y,y_3} \cong K_{x,x_3}$ is not isomorphic to K_{x,x_1} or K_{x,x_2}. By Claim 2, Spoiler can win in at most $b_0 + \log_2 h + O(1)$ extra moves, as required.

Case 2. In K there is a cycle C of length at most $2h+2$ containing the edges xx_1, xx_2 and a path of length at most h connecting y to a vertex of C via the edge yy_3.

By the assumption on h the vertex y cannot belong to a cycle of length at most $2h+2$ in K. Spoiler can point out this difference between x and y in at most $\log_2 h + O(1)$ moves (*cf.* the proof of Lemma 3.4). □

Claim 4. For the coloured graph K we have $D(K) \leqslant b_1 + O(1)$.

Proof of claim. Let $K' \ncong K$.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

For $x \in V(K)$, define a formula Φ_x with one free variable by

$$\Phi_x := \bigwedge_{y \in V(K) \setminus \{x\}} \Phi_{x,y} \tag{4.9}$$

with $\Phi_{x,y}$ as in Claim 3. Clearly, Φ_x has quantifier depth at most $b_1 + O(1)$.

If there is an $x \in V(K)$ with no $x' \in V(K')$ such that $K', x' \models \Phi_x$, then Spoiler selects x. Whatever Duplicator's reply x' is, it evaluates differently from x on Φ_x and Spoiler can win in at most $D(\Phi_x)$ further moves, as required. If there are two distinct $y', z' \in K'$ such that $K', y' \models \Phi_x$ and $K', z' \models \Phi_x$, then Spoiler selects both y' and z'. At least one of Duplicator's replies is not equal to x, say, $y \neq x$. Again, the selected vertices $y \in V(K)$ and $y' \in V(K')$ are distinguished by Φ_x, so Spoiler can win in at most $D(\Phi_x) + 2$ moves in total.

Therefore, let us assume that for every $x \in V(K)$ there is the unique vertex $x' = \phi(x) \in V(K')$ such that $K', x' \models \Phi_x$. Clearly, ϕ is injective. Furthermore, ϕ is surjective for if $x' \notin \phi(V(K))$, then Spoiler wins in $b_1 + O(1)$ moves: he selects $x' \in V(K')$ and then uses Φ_x, where $x \in V(K)$ is Duplicator's reply.

As $K \ncong K'$, Spoiler can select $x, y \in V(K)$ such that the adjacencies between x and y and between $x' = \phi(x)$ and $y' = \phi(y)$ are distinct or the vertex-colourings of $\{x, y\}$ and $\{x', y'\}$ are distinct. If Duplicator replies with x' and y', then she has lost. Otherwise she has violated ϕ and Spoiler wins in at most $b_1 + O(1)$ moves. ▯

The proof of Lemma 4.5 is complete by Claims 1 and 4. □

4.1.2. Probabilistic part. Here we estimate the parameters from the previous section. As before, let G be the giant component of $\mathcal{G}(n, \frac{c}{n})$ with $c > 1$, let C and K be its core and kernel, and let the parameters t, u, and d be defined by (4.5) and (4.6).

It is well known that w.h.p. $u = O(\frac{\ln n}{\ln \ln n})$ (see, *e.g.*, Bollobás [2, Chapter 3]), and $d = O(\ln n)$ (see, *e.g.*, Chung and Lu [3]).

Lemma 4.6. *With high probability every edge of K corresponds to at most $O(\ln n)$ vertices of G. Similarly, for any $x \in V(C)$ we have $v(T_x) = O(\ln n)$.*

Proof. The expected number of K-edges, each corresponding to precisely i vertices in G, is at most

$$f_i = \binom{n}{i}\binom{i}{2} p^{i-1} i^{i-2} (1-p)^{(i-2)(n-i)+\binom{i}{2}-i+1}.$$

If $i = o(n)$, then

$$f_i \leqslant \frac{n^i}{i!}\,\frac{i^2}{2}\, p^{i-1} i^{i-2} \mathrm{e}^{-(c+o(1))i} \leqslant n i^2 \left(\frac{\mathrm{e}c}{\mathrm{e}^{c+o(1)}}\right)^i.$$

We have $\mathrm{e}c < \mathrm{e}^c$ for $c > 1$, so if i is large enough, $i > M \ln n$ with $M = M(c)$, then $f_i < n^{-2}$. If $\alpha = i/n = \Omega(1)$, then using the estimate $\binom{n}{i} \leqslant (\mathrm{e}^{1-\alpha/2+o(1)} n/i)^i$ we get

$$f_i \leqslant \left((1+o(1))\,\frac{\mathrm{e}c}{\mathrm{e}^{\alpha/2+c((1-\alpha)+\alpha/2)}}\right)^i \leqslant \left((1+o(1))\,\frac{\mathrm{e}c}{\mathrm{e}^c}\right)^i < n^{-2}.$$

Thus $\sum_{i > M \ln n}^{n} f_i = o(1)$ and the claim follows from Markov's inequality.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Similarly, the expected number of vertices x with $v(T_x) = i > M \ln n$ can be bounded from above by

$$n\binom{n-1}{i-1} p^{i-1} i^{i-2} (1-p)^{(i-1)(n-i)+\binom{i}{2}-i+1} < n^{-2}. \qquad \square$$

In particular, Lemma 3.5 implies that w.h.p. $t = O(\frac{\ln n}{\ln \ln n})$.

Set $h = 2 \ln \ln n$. Thus $2^h / \ln n \to \infty$. It remains to prove that this choice of h satisfies Assumptions 1 and 2.

Lemma 4.7. *With high probability any set of $s \leqslant 6h + 5$ vertices of K spans at most s edges* (*including multiple edges and loops*).

Proof. A moment's thought reveals that it is enough to consider sets spanning connected subgraphs only.

Let $L = M \ln n$ be given by Lemma 4.6. The probability that there is a set S such that $|S| = s \leqslant 6h + 5$ and $K[S]$ is a connected graph with at least $s + 1$ edges is at most

$$\begin{aligned} o(1) + \sum_{s=1}^{6h+5} \binom{n}{s} s^{s-2} s^4 \sum_{0 \leqslant \ell_1, \dots, \ell_{s+1} \leqslant L} \prod_{i=1}^{s+1} \binom{n}{\ell_i} (\ell_i + 2)^{\ell_i} p^{\ell_i + 1} (1-p)^{\ell_i (n - \ell_i - 2)} \\ \leqslant o(1) + \sum_{s=1}^{6h+5} \left(\frac{n\mathrm{e}}{s}\right)^s s^{s+2} \sum_{0 \leqslant \ell_1, \dots, \ell_{s+1} \leqslant L} \prod_{i=1}^{s+1} \left(\frac{c\mathrm{e}^2}{n} \ell_i^2 \left(\frac{\mathrm{e}c}{\mathrm{e}^c} \right)^{\ell_i} \right) \\ \leqslant o(1) + \sum_{s=1}^{6h+5} \frac{(O(1))^s}{n} \left(\sum_{\ell=0}^{L} \ell^2 \left(\frac{\mathrm{e}c}{\mathrm{e}^c} \right)^{\ell} \right)^{s+1} \leqslant o(1) + \sum_{s=1}^{6h+5} \frac{(O(1))^s}{n} = o(1). \end{aligned}$$

The lemma is proved. □

Lemma 4.8. *With high probability K does not contain four vertices x, x', y, y' such that $\{x, y\}, \{x', y'\} \in E(K)$, $v(K_{x,y}) \geqslant 2^h$, $A_{x,y} \cap A_{x',y'} = \emptyset$, and $K_{x,y} \cong K_{x',y'}$.*

Proof. Let us briefly outline the proof. Let $H \in \mathcal{G}(n, p)$. For a pair of vertices $x, y \in V(H)$ we define a certain breadth-first search procedure $B(x, y)$ in H that resembles the construction of $K_{x,y}$ in K. Then we define an event FAIL(x, y, x', y') whose probability is $o(n^{-4})$. The proof will be complete when we show that if x, y, x', y', H do not satisfy the conclusion of the lemma, then FAIL(x, y, x', y') occurs. Please note that when we define $B(x, y)$ we do not assume that $\{x, y\} \in E(K)$; in fact, x and y may even lie outside the giant component.

Here are the details. Given c, choose the following small positive constants in this order: $\varepsilon_1 \gg 1/M_1 \gg \varepsilon_2 \gg 1/M_2 \gg \epsilon_3$. Let $x, y, x', y' \in V(H)$. Next, we describe the procedures $B(x, y)$ and $B(x', y')$, and specify which outcomes are included in FAIL(x, y, x', y').

In the procedure $B(x, y)$ we take the breadth-first search in $H - x$ starting with y. Let $L_1 = \{y\}$, L_2, L_3, *etc.*, be the levels. Let $T_i = \cup_{j=1}^{i} L_i$. Let s be the smallest index such that $|T_s| \geqslant M_1 \ln n$. If the search dies out before we reach $M_1 \ln n$ vertices, then s and the parameters depending on it are undefined but this itself does not result in FAIL.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

If $|T_s| > 2cM_1 \ln n$, then this is FAIL. Chernoff's bound implies that the probability of this event is $o(n^{-4})$. Indeed, this is at most the probability that the binomial random variable with parameters $(n, \frac{c}{n} \times M_1 \ln n)$ exceeds $2cM_1 \ln n$.

We also get FAIL if there is an $i \geqslant s$ such that $|T_i| \leqslant \epsilon_3 n$ and $|L_{i+1}|$ does not lie inside the interval $(c \pm \varepsilon_2)|L_i|$. Again, by Chernoff's bound this has probability $o(n^{-4})$. Informally speaking, the levels L_i increase proportionally after we reached $M_1 \ln n$ vertices.

Take some $i \geqslant s$ with $|T_i| \leqslant \epsilon_3 n$. The sizes of the next M_2 levels of the breadth-first search from the vertices of L_i can be bounded from below by *independent* branching processes with the number of children having the Poisson distribution with mean $c - \varepsilon_2$. Indeed, for every active vertex v choose a pool P of $\lceil (1 - \frac{\varepsilon_2}{2c})n \rceil$ available vertices and let v choose its neighbours from P, each with probability c/n. If v claimed r neighbours, then, when we take the next active vertex, we add extra r vertices to the pool, so that its size remains constant.

With positive probability p_1 the ideal branching process survives infinitely long; in fact, p_1 is the positive root of $1 - p_1 = \mathrm{e}^{-(c-\epsilon_2)p_1}$. Let

$$p_2 = \max_{j \geqslant 0} \frac{c^j \mathrm{e}^{-c}}{j!} < 1.$$

The numbers $p_1 > 0$ and $p_2 < 1$ are constants (depending on c only).

Take the smallest q such that $|T_q| \geqslant M_2 \ln n$. We know that

$$|L_q| \geqslant \left(\frac{c-1}{c} - \varepsilon_1 \right) |T_q| \tag{4.10}$$

because the levels grow proportionally from the sth level (and $|T_s| \leqslant 2cM_1 \ln n \ll M_2 \ln n \leqslant |T_q|$). Let Z consist of those vertices of L_q for which the search process goes on for at least M_2 further levels before dying out.

We define $B(x', y')$ in the same way, using the same notation but with primes added (for example, L'_i, T'_i, and so on). Now we specify the last (and crucial) component of the event FAIL. It depends on all four vertices x, y, x', y'.

We fail if s is defined, $T_{q+M_2} \cap T'_{q+M_2} = \emptyset$, and there is an isomorphism $\phi : H[T_{q+M_2}] \cong H[T'_{q+M_2}]$ such that $\phi(y) = y'$. Let us analyse this event. We expose T_{q+M_2} and all edges inside it. The graph $H' = H - T_{q+M_2}$ is the genuine Erdős–Rényi random graph with edge probability p. Also, we know that $|T_{q+M_2}| = o(n)$. Assume $T_{q+M_2} \cap T'_{q+M_2} = \emptyset$. This means that T'_{q+M_2-1} can be alternatively constructed by applying the procedure $B(x', y')$ to the random graph H'.

The expected number of embeddings of $\phi : H[T_q] \to H'$ respecting the special vertices y and y' is at most $n^{|T_q|-1} p^{|T_q|-1}$. If FAIL occurs, then on top of $\phi(H[T_q])$, we have to get the $|Z|$ specified trees, each of height at least M_2. But the probability of this event is at most

$$\left((p_2 + o(1))^{M_2} \right)^{|Z|}, \tag{4.11}$$

because if we want to get a given height-M_2 tree, then at least M_2 times we have to match the sum of degrees of a level, each coincidence having probability at most $p_2 + o(1)$.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Our previous assumptions and (4.10) imply that with probability $1 - o(n^{-4})$

$$|Z| \geqslant \frac{p_1}{2} \times |L_q| \geqslant \frac{p_1}{2} \times \left(\frac{c-1}{c} - \epsilon_1\right) \times |T_q|.$$

By (4.11) the probability of failure at this stage is at most

$$n^{|T_q|-1}\, p^{|T_q|-1}(p_2 + o(1))^{M_2 \, \frac{p_1}{2}\, \left(\frac{c-1}{c} - \epsilon_1\right) |T_q|},$$

which is $o(n^{-4})$ because $|T_q| \geqslant M_2 \ln n$ and the constant M_2 can be arbitrarily large.

Claim 1. Suppose that x, y, x', y', H do not satisfy the assumptions of the lemma. Then we have $\mathrm{FAIL}(x, y, x', y')$.

Proof of claim. Since $v(K_{x,y}) \geqslant 2^h > M_2 \ln n$, the parameters s, q, and others are defined. Also, $T_{q+M_2} \subset V(G_{x,y})$ and $T'_{q+M_2} \subset V(G_{x',y'})$, so their intersection is empty. Since $K_{x,y} \cong K_{x',y'}$, we have $G_{x,y} \cong G_{x',y'}$ and $H[T_{q+M_2}] \cong H[T'_{q+M_2}]$. Thus the event $\mathrm{FAIL}(x, y, x', y')$ occurs, as required. □

Now we are ready to complete the proof of the lemma. There are at most n^4 choices for x, y, x', y' and each violates the conclusion of the lemma with probability $o(n^{-4})$ by Claim 1. By Markov's inequality, w.h.p. no violations happen. □

Putting all together we deduce the upper bound of Theorem 4.1.

4.2. Lower bound

Let $l = (1-\varepsilon)\frac{\ln n}{\ln \ln n}$ for some $\varepsilon > 0$. We claim that w.h.p. the giant component G has a vertex i adjacent to at least l leaves of G. (Then we have $D(G) \geqslant l + 1$: consider the graph obtained from G by adding an extra leaf attached to i.)

Choose a constant $\delta > 0$ so that $c(1-\delta) > 1$ and let V' be a fixed set of $\lfloor (1-\delta)n \rfloor$ vertices of $V(H)$, where $H \in \mathcal{G}(n, p)$. First we expose $H' = H[V']$. The expected degree in H' is $p(|V'| - 1) > 1$, so we are in the supercritical stage and w.h.p. there is the giant component G' of order $m = \Omega(n)$.

Let us expose the remaining edges of H. For $i \in V(G')$ let X_i be the event that, in H, the vertex i is incident to at least l leaves from $D = V \setminus V'$. It is easy to estimate the expectation of $X = \sum_{i \in V(G')} X_i$:

$$\begin{aligned} E(X) &= m\binom{|D|}{l} p^l (1-p)^{\binom{l}{2} + l(n-l-1)} + O(1) \times m\binom{|D|}{l+1} p^{l+1} (1-p)^{(l+1)n} \\ &= (1 + o(1)) \frac{m\delta^l c^l \mathrm{e}^{-cl}}{l!} \;\to\; \infty. \end{aligned}$$

Also, for $i \neq j$,

$$\begin{aligned} E(X_i \wedge X_j) &= (1 + o(1)) \binom{|D|}{l} \binom{|D| - l}{l} p^{2l} (1-p)^{\binom{2l}{2} + 2l(n-2l-1)} \\ &= (1 + o(1))\, E(X_i) E(X_j). \end{aligned}$$

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

The second moment method gives that X is concentrated around its mean. In particular, w.h.p. $X > 0$. Since every vertex of G' also belongs to the giant component of H, the claim follows.

5. Random trees

We consider the probabilistic model $\mathscr{T}(n)$, where a tree T on the vertex set $[n]$ is selected uniformly at random among all n^{n-2} trees. In this section we prove that w.h.p. $D(T)$ is close to the maximum degree of T.

Theorem 5.1. *Let $T \in \mathscr{T}(n)$. With high probability $D(T) = (1+o(1))\Delta(T) = (1+o(1))\frac{\ln n}{\ln\ln n}$.*

Let $\mathscr{F}(n,k)$ be a forest chosen uniformly at random from the family of $\mathscr{F}_{n,k}$ of all forests with the vertex set $[n]$, which consist of k trees rooted at vertices $1,2,\ldots,k$. Note that a random tree $T \in \mathscr{T}(n)$ can be identified with $\mathscr{F}(n,1)$. We recall that $|\mathscr{F}_{n,k}| = kn^{n-k-1}$; see, *e.g.*, Stanley [13, Theorem 5.3.2]. We start with the following simple facts on $\mathscr{F}(n,k)$.

Lemma 5.2. *Let $k = k(n) = o(\sqrt{n})$.*

(i) *The probability that $\mathscr{F}(n,k)$ contains precisely ℓ, $0 \leqslant \ell \leqslant k-1$, isolated vertices is* $(1+O(k^2/n))\binom{k-1}{\ell}\mathrm{e}^{-\ell}(1-\mathrm{e}^{-1})^{k-\ell-1}$.

(ii) *Let $k \leqslant \ln^4 n$ and $k_0 = k(1+1/\ln n) + 9\ln^2 n$. The probability that the roots of $\mathscr{F}(n,k)$ have more than k_0 neighbours combined is $o(n^{-3})$.*

(iii) *The probability that ℓ given roots of $\mathscr{F}(n,k)$ have degree at least $s \geqslant 4$ each is bounded from above by $(2/(s-1)!)^\ell$*

Proof. In order to see (i), note that from the generalized inclusion–exclusion principle the stated probability equals

$$\begin{aligned}\sum_{i=\ell}^{k-1}\binom{i}{\ell}(-1)^{i-\ell}\binom{k}{i}\frac{(k-i)(n-i)^{n-k-1}}{kn^{n-k-1}}\\ &= \left(1+O\left(\frac{k^2}{n}\right)\right)\sum_{i=\ell}^{k-1}\frac{(k-1)!}{\ell!(i-\ell)!(k-1-i)!}(-1)^{i-\ell}\mathrm{e}^{-i}\\ &= \left(1+O\left(\frac{k^2}{n}\right)\right)\binom{k-1}{\ell}\mathrm{e}^{-\ell}(1-\mathrm{e}^{-1})^{k-\ell-1}. \qquad (5.1)\end{aligned}$$

Let us prove (ii). For the probability that precisely m vertices of $\mathscr{F}(n,k)$ are adjacent to the roots, where $k \leqslant m \leqslant n-k$, Stirling's formula gives, rather crudely,

$$\binom{n-k}{m}k^m\frac{m\,(n-k)^{n-k-m-1}}{k\,n^{n-k-1}} \leqslant O(1)\left(\frac{\mathrm{e}^{1-k/m}k}{m}\right)^m\frac{\sqrt{(n-k)m}}{k}. \qquad (5.2)$$

For every x, $0 < x < 1$, we have $x\mathrm{e}^{1-x} \leqslant \mathrm{e}^{-(1-x)^2/2}$. Thus, for $m > k_0$,

$$\left(\frac{\mathrm{e}^{1-k/m}k}{m}\right)^m \leqslant \exp\left(-\frac{(m-k)^2}{2m}\right) \leqslant \begin{cases}\exp\left(-\frac{(9\ln^2 n)^2}{18\ln^3 n}\right), & \text{if } m \leqslant 9\ln^3 n,\\ \exp\left(-\frac{(m/(1+\ln n))^2}{2m}\right), & \text{if } m > 9\ln^3 n.\end{cases}$$

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

This is at most $e^{-4.4\ln n} = n^{-4.4}$ in either case. Thus

$$\sum_{m>k_0} \exp\left(-\frac{(m-k)^2}{2m}\right) \leqslant n \times n^{-4.4} = o(n^{-3}),$$

proving the assertion.

For $k=1$ the probability that a given root has degree at least s is bounded from above by

$$\sum_{t\geqslant s} \binom{n-1}{t} \frac{t(n-1)^{n-t-2}}{n^{n-2}} \leqslant \sum_{t\geqslant s} \frac{1}{(t-1)!} \leqslant \frac{2}{(s-1)!}.$$

If we fix some $\ell \geqslant 2$ roots, then if we condition on the vertex sets of the ℓ corresponding components, the obtained trees are independent and uniformly distributed, implying the required bound by the above calculation. □

Using the above result one can estimate the number of vertices of $T \in \mathscr{T}(n)$ with a prescribed number of pendant neighbours.

Lemma 5.3. *Let $X_{\ell,m}$ denote the number of vertices in $T \in \mathscr{T}(n)$ with precisely ℓ neighbours of degree one and m neighbours of degree larger than one. Let*

$$A \subseteq \{(\ell,m):\ 0 \leqslant \ell \leqslant \ln n, \quad 1 \leqslant m \leqslant \ln n\},$$

be a set of pairs of natural numbers and $X_A = \sum_{(\ell,m)\in A} X_{\ell,m}$. Then, the expectation

$$E(X_A) = (1+o(1))\, n \sum_{(\ell,m)\in A} \frac{e^{-\ell-1}}{\ell!} \frac{(1-e^{-1})^{m-1}}{(m-1)!} \tag{5.3}$$

and, if $E(X_A)/\ln^3 n \to \infty$, then

$$E(X_A(X_A-1)) = (1+o(1))\,(E(X_A))^2. \tag{5.4}$$

Proof. Using Lemma 5.2(i) we get

$$E(X_A) = (1+o(1))n \sum_{(\ell,m)\in A} \binom{n-1}{m+\ell} \frac{(m+\ell)(n-1)^{n-m-\ell-2}}{n^{n-2}} \times \binom{m+\ell-1}{\ell} e^{-\ell}(1-e^{-1})^{m-1},$$

which gives (5.3).

Let $E(X_A)/\ln^3 n \to \infty$. In order to count the expected number of pairs of vertices with prescribed neighbourhoods, we condition on the event that a vertex x has $\ell+m$ neighbours $x_1,\dots,x_\ell,y_1,\dots,y_m$ such that $d(x_i)=1$ and $d(y_i)\geqslant 2$. The graph $F = T - \{x,x_1,\dots,x_\ell\}$ is a forest with roots $y_1,\dots,y_m$. Once we condition on the orders of the components $T_1,\dots,T_m$ of F, say $n_1+\dots+n_m = n-\ell-1$ with each $n_i \geqslant 2$, each component is distributed as a random tree. For each T_i we estimate the expectation of the corresponding random variable $X_{A,i}$ using the same argument as above. If $n_i/\ln^2 n \to \infty$, then the changes are negligible and the asymptotic estimate (5.3) still applies to $X_{A,i}$. Note that the error term

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

$o(1)$ in (5.3) is in fact $O(\max\{(\ell+m)^2 : (\ell,m) \in A\}/n)$ because of the error term $O(k^2/n)$ in Lemma 5.2(i). The other components contain at most $O(\ln^3 n)$ vertices: recall that $m \leqslant \ln n$ for any $(\ell,m) \in A$. Hence, $\sum_{i=1}^{m} E(X_{A,i}) = (1+o(1))\, E(X_A)$, and (5.4) holds. □

As an easy corollary of the above result we get a lower bound for $D(\mathscr{T}(n))$.

Theorem 5.4. *Let $T \in \mathscr{T}(n)$. With high probability $D(T) \geqslant (1-o(1))\Delta(T) = (1-o(1))\,\frac{\ln n}{\ln\ln n}$.*

Proof. Since w.h.p. the maximum degree is

$$\Delta(T) = (1+o(1))\,\frac{\ln n}{\ln\ln n} \tag{5.5}$$

(see Moon [6]), in order to prove the assertion it is enough to show that w.h.p. T contains a vertex v with

$$\ell_0 = (1-o(1))\,\frac{\ln n}{\ln\ln n} \tag{5.6}$$

neighbours of degree one. Indeed, to characterize such a structure Spoiler needs at least $\ell_0 + 1$ moves. Using Lemma 5.3, we infer that, for the number of vertices X_ℓ of T with exactly ℓ neighbours of degree 1 and, say, one neighbour of degree larger than 1, we have $E(X_\ell) = \Omega(\mathrm{e}^{-\ell} n/\ell!)$. Thus, one can choose ℓ_0 so that (5.6) holds and $E(X_{\ell_0})/\ln^3 n \to \infty$. Then, due to Lemma 5.3, $\mathrm{Var}(X_{\ell_0}) = o((E(X_{\ell_0}))^2)$, and Chebyshev's inequality implies that w.h.p. $X_{\ell_0} > 0$. □

Let us state another simple consequence of Lemma 5.2 which will be used in our proof of Theorem 5.1. Here and below $N_r(v)$ denotes the r-neighbourhood of v, *i.e.*, the set of all vertices of the graph which are at distance r from v, and $N_{\leqslant r}(v) = \bigcup_{i=0}^{r} N_i(v)$.

Lemma 5.5. *Let $r_0 = r_0(n) = \lceil 7\ln n \rceil$. Then, w.h.p. the following holds for every vertex v of $T \in \mathscr{T}(n)$:*

(i) $|N_{\leqslant r_0}(v)| \leqslant 10^8 \ln^4 n$,

(ii) $N_{\leqslant r_0}(v)$ *contains fewer than* $\ln n/(\ln\ln n)^2$ *vertices of degree larger than* $(\ln\ln n)^5$.

Proof. For $s \leqslant r_0$ let $W_s = N_{\leqslant s}(v)$. Note that, conditioned on the structure of the subtree of T induced by W_s for some $s \leqslant r_0$, the forest $T - W_{s-1}$ can be identified with the random forest on $n - |W_{s-1}|$ vertices, rooted at the set $N_s(v)$. Thus, it follows from Lemma 5.2(ii) that w.h.p. the following holds: let i be the smallest integer with $|N_i(v)| \geqslant 9\ln^3 n$; then $|N_i(v)| \leqslant 10\ln^3 n$ and for every $j \geqslant i$ with $|N_j(v)| \leqslant \ln^4 n$ we have $|N_{j+1}(v)| \leqslant |N_j(v)|(1 + 2/\ln n)$. Hence,

$$|N_{\leqslant r_0}(v)| \leqslant r_0 \times 10\ln^3 n \times (1+2/\ln n)^{r_0} \leqslant 10^8 \ln^4 n.$$

In order to show (ii) note that (i) and Lemma 5.2(iii) imply that the probability that, for some vertex v, at least $\ell = \lfloor \ln n/(\ln\ln n)^2 \rfloor$ vertices of $N_{\leqslant r_0}(v)$ have degree larger than

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

$m = (\ln\ln n)^5$ is bounded from above by

$$n\binom{\ln^5 n}{\ell}\left(\frac{2}{(m-1)!}\right)^{\ell} \leqslant n\left(\frac{2\mathrm{e}\ln^5 n}{\ell(m-1)!}\right)^{\ell} = o(1). \qquad \square$$

In our further argument we need some more definitions. Let T be a tree and let v be a vertex of T. For a vertex $w \in N_r(v)$ let P_{vw} denote the unique path connecting v to w (of length r). Let the *check* $\mathrm{Ch}(v, P_{vw})$ be the binary sequence $b_0 \cdots b_r$, in which, for $i = 0, \ldots, r$, b_i is zero (resp. 1) if the ith vertex of P_{vw} is adjacent (resp. not adjacent) to a vertex of degree one. (The parameter v in $\mathrm{Ch}(v, P_{vw})$ is needed to indicate in which direction we go along P_{vw}.) Finally, the r-*checkbook* $\mathrm{Ch}_r(v)$ is the set

$$\mathrm{Ch}_r(v) = \{\mathrm{Ch}(v, P_{vw}) : w \in N_r(v)\}.$$

Note that a checkbook is not a multiset, *i.e.*, a check from $\mathrm{Ch}_r(v)$ may correspond to more than one of the paths P_{vw}.

Our proof of the upper bound for $D(\mathscr{T}(n))$ is based on the following fact.

Theorem 5.6. *Let $r_0 = \lceil 7\ln n\rceil$. With high probability, for each pair P_{vw}, $P_{v'w'}$ of paths of length r_0 in $T \in \mathscr{T}(n)$ which share at most one vertex, the checks $Ch(v, P_{vw})$ and $Ch(v', P_{v'w'})$ are different.*

Proof. Let $C = \mathrm{del}(T)$ denote the tree obtained from T by removing all vertices of degree one. From Lemma 5.3 it follows that w.h.p. the tree C has $(1 - \mathrm{e}^{-1} - o(1))n$ vertices, of which

$$(1 + o(1))\, n \sum_{\ell > 0} \frac{\mathrm{e}^{-\ell-1}}{\ell!} = (\exp(\mathrm{e}^{-1} - 1) - \mathrm{e}^{-1} + o(1))\, n$$

vertices have degree one, and

$$\alpha n = (1 - \exp(\mathrm{e}^{-1} - 1) + o(1))\, n$$

vertices have degree greater than one. (Note that by (5.5) there is no need to consider the values of ℓ larger than $\ln n$ in Lemma 5.3.)

Moreover, among the set B of $(\mathrm{e}^{-1} + o(1))n$ vertices removed from T,

$$(1 + o(1))n \sum_{l=1}^{\infty} \ell \frac{\mathrm{e}^{-\ell-1}}{\ell!} = (1 + o(1)) \exp(\mathrm{e}^{-1} - 2)n$$

were adjacent to vertices which became pendant in C. Let B' denote the set of the remaining

$$(\mathrm{e}^{-1} - \exp(\mathrm{e}^{-1} - 2) + o(1))n = (\rho_0 + o(1))n$$

vertices which are adjacent to vertices of degree at least two in C. Note that, given $C = \mathrm{del}(T)$, each attachment of vertices from $B \setminus B'$ to pendant vertices of C, such that each pendant vertex of C gets at least one vertex from $B \setminus B'$, as well as each attachment of vertices from B' to vertices of degree at least two from C, is equally likely.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Let P_{vw}, $P_{v'w'}$, be two paths of length r_0 in T which share at most one vertex. Clearly, each vertex of P_{vw}, except at most two vertices at each of the ends, belongs to C and contains at least two neighbours in C; the same is true for $P_{v'w'}$. Since $(\rho_0 + o(1))n$ vertices from B' are attached to the αn vertices of degree at least two in C at random, the probability that one such vertex gets no attachment is

$$p_0 = (1 + o(1))\left(1 - \frac{1}{\alpha n}\right)^{\rho_0 n} = (1 + o(1))\,\mathrm{e}^{-\rho_0/\alpha} = 0.692\ldots + o(1).$$

Therefore, the probability that the checks $\mathrm{Ch}(v, P_{vw})$ and $\mathrm{Ch}(v', P_{v'w'})$ are identical can be bounded from above by

$$o(n^{-3}) + \left(p_0^2 + (1 - p_0)^2 + o(1)\right)^{r_0} \leqslant \mathrm{e}^{-3\ln n} = o(n^{-2}\ln^{-8} n).$$

Indeed, Chernoff's bound implies that with probability $1 - o(n^{-4})$, for every path P in T of length r_0 the number of leaves that attach to the path is $o(n)$. So, if we expose one by one the attachments to pairs of the corresponding vertices of P_{vw} and $P_{v'w'}$ and each time condition on $o(n)$ exposed leaves so far, this would change the total probability by at most $r_0 n^{-4} = o(n^{-3})$. But the conditional probability that a given pair of distinct vertices yields a coincidence is clearly $p_0^2 + (1 - p_0)^2 + o(1)$.

Since, by Lemma 5.5(i), w.h.p. T contains at most $O(n\ln^4 n)$ checks of length r_0, the assertion follows. □

Now, let $r_0 = \lceil 7\ln n\rceil$. We call a tree T on n vertices *typical* if:

- for each pair of paths P_{vw}, $P_{v'w'}$ of length r_0 which share at most one vertex, the checks $\mathrm{Ch}(v, P_{vw})$, $\mathrm{Ch}(v', P_{v'w'})$ are different,
- for the maximum degree Δ of T we have
$$\frac{\ln n}{2\ln\ln n} \leqslant \Delta \leqslant \frac{2\ln n}{\ln\ln n},$$
- $|N_{\leqslant r_0}(v)| \leqslant 10^8 \ln^4 n$, for every vertex v,
- for every vertex v at most $\ln n/(\ln\ln n)^2$ vertices of degree larger than $(\ln\ln n)^5$ lie within distance r_0 from v.

Theorem 5.7. *For a typical tree $T \in \mathscr{T}(n)$ we have $D(T) \leqslant (1 + o(1))\,\Delta$.*

Proof. Our approach is somewhat similar to that for the giant component from Section 4.

Let T be a typical tree and let T' be any other graph which is not isomorphic to T. We shall show that Spoiler can win the Ehrenfeucht game on T and T' in $(1 + o(1))\Delta$ moves.

Let us call a vertex v of a graph a *yuppie* if there are two paths P_{vw}, $P_{vw'}$ of length r_0 starting at v so that $V(P_{vw}) \cap V(P_{vw'}) = \{v\}$. Note that the set of all yuppies Y spans a subtree in T, call it K.

Claim 1. Every vertex v is at distance at most r_0 from a yuppie.

Proof of claim. By the assumption on the neighbourhoods, we have

$$|N_{<2r_0}(v)| \leqslant 10^{16}\ln^8 n < n,$$

so there is a vertex u at distance $2r_0$ from v. The median of the vu-path is a yuppie. ▯

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Let us view K as a vertex-coloured graph where the colour of a vertex $x \in Y$ is the isomorphism type of T_x, the component of $T - (Y \setminus \{x\})$ rooted at x. Let Y' be the set of yuppies of T', and let $K' = T'[Y']$. We can assume that Duplicator respects the subgraphs K and K', for otherwise Spoiler wins in extra $O(\ln \ln n)$ moves.

Claim 2. Any distinct $v, v' \in K$ can be distinguished in $O(\ln \ln n)$ moves.

Proof of claim. Assume that the r_0-checkbooks of v, v' are the same, for otherwise Spoiler wins in $\log_2 r_0 + O(1)$ moves. (Please note that the checkbooks are viewed as sets, not as multisets, so the number of moves does not depend on the degrees of v and v'.)

Take a path P_{vx} of length r_0, which shares with $P_{vv'}$ only vertex v. Spoiler selects x. Let Duplicator reply with x'. Assume that $\mathrm{Ch}(v, P_{vx}) = \mathrm{Ch}(v', P_{v'x'})$. The path $P_{v'x'}$ must intersect P_{vx}; thus $v \in P_{v'x'}$. Next, Spoiler selects the P_{vx}-neighbour y of v. Assume that Duplicator's reply y' belongs to $P_{v'x'}$ (for otherwise Spoiler can win in $O(\ln \ln n)$ moves).

Let $z \in T$ maximize $\mathrm{dist}(v, z)$ on the condition that $\mathrm{Ch}_{r_0}(z) = \mathrm{Ch}_{r_0}(v)$ and v lies between y and z in T. Define the analogous vertex z', replacing v, y in the definition by v', y'. We have $\mathrm{dist}(v, z) > \mathrm{dist}(v', z')$ because any legitimate choice for z' is also a legitimate choice for z. Let Spoiler select $w = z$. If Duplicator's reply w' satisfies $\mathrm{Ch}_{r_0}(w') \neq \mathrm{Ch}_{r_0}(w)$, then Spoiler wins in at most $\log_2 r_0 + O(1)$ extra moves. Otherwise, v' is not on the path between y' and w' or $\mathrm{dist}(v, w) > \mathrm{dist}(v', w')$. Moreover, $\mathrm{dist}(v, w) \leqslant 2r_0$ (because their r_0-checkbooks are equal). Spoiler wins in $\log_2 r_0 + O(1)$ extra moves. The claim has been proved. □

Similarly to the argument following (4.9), one can argue that, for every vertex $x \in K$, there is a formula Φ_x with a single free variable of quantifier depth $O(\ln \ln n)$ identifying x in T (note that the property of being a yuppie is definable with depth $\log_2 r_0 + O(1)$). Moreover, we can assume that this gives us an isomorphism $\phi : K \to K'$ which is respected by Duplicator.

Assume first that T' is not connected. As $K' \cong K$ is connected, there is a component C' of T' without a yuppie. Spoiler chooses an $x' \in C'$. Now, any Duplicator's reply x is within distance r_0 from a yuppie by Claim 1, which is not true for x'. Spoiler can win in $O(\ln \ln n)$ moves.

Assume now that T' is connected. It follows that there is a vertex $x \in K$ such that $T_x \ncong T'_{x'}$, where $x' = \phi(x)$ and $T'_{x'}$ is the component of $T' - (Y' \setminus \{x'\})$ rooted at x'.

Since each vertex of T is within distance at most r_0 from some yuppie by Claim 1, the tree T_x has height at most r_0. If $T'_{x'}$ has a path of length greater than $2r_0$ or a cycle, then Spoiler can win in at most $\log_2 r_0 + O(1)$ moves; *cf.* Lemma 3.4. (Duplicator is forced to play within T_x and $T'_{x'}$ on account of Lemma 3.3, because Y and Y' are succinctly definable.) So assume that $T'_{x'}$ is a tree. Now Spoiler should select all vertices of T_x which are of degree larger than $(\ln \ln n)^5$, say $w_1, \dots, w_s$. Since T is typical there are at most $\ln n / (\ln \ln n)^2$ such vertices in T_x. Suppose that, in response, Duplicator chooses vertices $w'_1, \dots, w'_s$ in $T'_{x'}$. Then, $T_x \setminus \{w_1, \dots, w_s\}$ splits into a number of trees $F_1, \dots, F_u$, coloured according to their adjacencies to the w_i's. Now, for some i the multisets of coloured trees adjacent to w_i and w'_i are different. Spoiler can highlight this by using at most

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

$\Delta(T)+1$ moves. Now Spoiler plays inside some F_i, using the strategy of Theorem 3.1. Note that F_i has at most $10^8 \ln^4 n$ vertices and maximum degree at most $(\ln\ln n)^5$, so $O((\ln\ln n)^6/\ln\ln\ln n)$ moves suffice here.

Consequently, for a typical tree T,

$$D(T) \leqslant \Delta(T) + \frac{\ln n}{(\ln\ln n)^2} + o((\ln\ln n)^6),$$

and the assertion follows. □

Proof of Theorem 5.1. Theorem 5.1 is an immediate consequence of Theorems 5.4 and 5.7 and the fact that, due to Lemma 5.5, Theorem 5.6, and the known estimates of maximum degree [6], w.h.p. a random tree $T \in \mathscr{T}(n)$ is typical. □

6. Restricting alternations

If Spoiler can win the Ehrenfeucht game, alternating between the graphs G and G' at most r times, then the corresponding sentence has *alternation number* at most r, that is, any chain of nested quantifiers has at most r changes between $\exists$ and $\forall$. (To make this well defined, we assume that no quantifier is within the range of a negation sign.) Let $D_r(G)$ be the smallest quantifier depth of a sentence which defines G and has alternation number at most r. It is not hard to see that $D_r(G) = \max\{D_r(G,G') : G' \not\cong G\}$, where $D_r(G,G')$ may be defined as the smallest k such that Spoiler can win $\textsc{Ehr}_k(G,G')$ with at most r alternations. For small r, this is a considerable restriction on the structure of the corresponding formulas, so let us investigate the alternation number given by our strategies.

Let $D_r^{\mathrm{tree}}(n,l)$ be the maximum of $D_r(T)$ over all coloured trees of order at most n and maximum degree at most l.

Unfortunately, in Theorem 3.1 we have hardly any control on the number of alternations. However, we can show that alternation number 0 suffices if we are happy to increase the upper bound by a factor of 2.

Lemma 6.1. *Let T and T' be non-isomorphic coloured trees. Assume that $v(T) \geqslant v(T')$ and denote $n = v(T)$. Assume also that $\Delta(T) \leqslant l$ and let both l and $\ln n/\ln l$ tend to infinity. Then Spoiler can win the Ehrenfeucht game on (T,T') in at most*

$$(1+o(1))\frac{l \ln n}{\ln l} \tag{6.1}$$

moves, playing all the time in T.

Proof. In the first move Spoiler selects a median $x \in T$; let x' be Duplicator's reply. If $d(x) > d(x')$, then Spoiler wins in extra l moves, which is negligible when compared to (6.1). So, suppose that $d(x') \geqslant d(x)$.

Let $t = d(x)$ and $C_1,\ldots,C_t$ be the (rooted) components of $T - x$ indexed so that $v(C_1) \geqslant v(C_2) \geqslant \cdots \geqslant v(C_t)$. Referring to the root of a component, we mean that vertex of it which is adjacent to x. Spoiler starts selecting, one by one, the roots of $C_1, C_2, \ldots, C_t$ in this order. Duplicator is forced to respond with roots of distinct components of $T' - x'$.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Spoiler keeps doing so until the following situation occurs: he selects the root y of a component $C = C_i$ while Duplicator selects the root y' of a component C' such that $v(C) \geqslant v(C')$ and $C \not\cong C'$ (as rooted trees). Such a situation must occur for some $i \leqslant t$ due to the conditions that $v(T) \geqslant v(T')$, $d(x) \leqslant d(x')$, and $T \not\cong T'$.

We claim that if Spoiler selects a vertex z inside C then Duplicator must reply with some $z' \in C'$, for otherwise Spoiler wins in at most $\log_2 n$ extra moves. Indeed, suppose $z' \notin C'$. Spoiler selects z_1, which is a middle point of the yz-path. Whatever the reply z_1', the $z'z_1'$-path or $z_1'y'$-path contains the vertex x'. Suppose it is the $z'z_1'$-path. Then Spoiler halves the zz_1-path, and so on. In at most $\log_2 n$ times he wins.

Thus, making $i+1 \leqslant t+1 \leqslant l+1$ steps, we have reduced the game to two non-isomorphic (rooted) trees, C and C', with $v(C) \leqslant \min(\frac{1}{i}, \frac{1}{2})\, v(T)$. In the game on (C, C') Spoiler applies the same strategy recursively. Two ending conditions are possible: the root of C has strictly larger degree than the root of C' and Duplicator violates a colour, the adjacency, or the equality relation. It is easy to argue (*cf.* the proof of Lemma 3.5) that the worst case for us is when we have $i = (1 + o(1))\, l$ all the time, which gives the required bound (6.1). □

Theorem 6.2. *Let both l and $\ln n/\ln l$ tend to infinity. Then*

$$D_0^{\mathrm{tree}}(n, l) \leqslant (1 + o(1)) \frac{l \ln n}{\ln l}. \tag{6.2}$$

Proof. Let T be a tree of order n and maximum degree at most l and let $G \not\cong T$. If $\Delta(T) \neq \Delta(G)$ then Spoiler wins the Ehrenfeucht game on (T, G) in at most $l+2$ moves playing in the graph of the larger degree. We therefore assume that T and G have the same maximum degree not exceeding l.

Case 1. G contains a cycle of length no more than $n+1$.

Spoiler plays in G proceeding as in the last paragraph of the proof of Lemma 3.4.

Case 2. G is connected and has no cycle of length up to $n+1$.

If $v(G) \leqslant n$, then G must be a tree. Lemma 6.1 applies. Let us assume $v(G) > n$. Let A be a set of $n+1$ vertices spanning a connected subgraph in G. This subgraph must be a tree. Spoiler plays in G, staying all the time within A. Lemma 6.1 applies.

Case 3. G is disconnected and has no cycle of length up to $n+1$.

We can assume that every component H of G is a tree, for otherwise Spoiler plays the game on (T, H) staying in H, using the strategy described above.

Suppose first that G has a tree component H such that $H \not\cong T$ and $v(H) \geqslant n$. If $v(H) = n$, let $T' = H$. Otherwise let T' be a subtree of H on $n+1$ vertices. Spoiler plays the game on (T, T') staying in T' and applying the strategy of Lemma 6.1 (with T and T' interchanged and perhaps with $n+1$ in place of n).

Suppose next that all components of G are trees of order less than n. In the first move Spoiler selects a median x of T. Let Duplicator respond with a vertex x' in a component T' of G. If in the sequel Duplicator makes a move outside of T', then Spoiler wins by Lemma 3.3. As long as Duplicator stays in T', Spoiler follows the strategy of Lemma 6.1.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press

Finally, it remains to consider the case that G has a component T' isomorphic to T. Spoiler plays in G. In the first move he selects a vertex x' outside T'. Let x denote Duplicator's response in T. Starting from the second move Spoiler plays the game on (T, T') according to Lemma 6.1, where x is considered coloured in a colour absent in T'.

Our description of Spoiler's strategy is complete. □

It is not clear what the asymptotic of $D_0^{\text{tree}}(n,l)$ is. We could not even rule out the possibility that $D_0^{\text{tree}}(n,l) = (\frac{1}{2} + o(1))\,\frac{l \ln n}{\ln l}$.

The similar method shows that $D_0^{\text{tree}}(n,l) = \Theta(\ln n)$ if $l \geqslant 2$ is constant and $D_0^{\text{tree}}(n,l) = \Theta(l)$ if $\frac{\ln n}{\ln l} = O(1)$, but the exact asymptotic seems difficult to compute.

Using these results, one can rewrite the proofs of Theorems 4.1 and 5.1 so that the obtained sentences have a small number of alternations (at most 3). However, we could not find strategies requiring no alternations at all. For example, one of a few places that seems to require an alternation is establishing that ϕ is a bijection: Spoiler may be forced to start in one of the graphs, while later (for example, when showing that $T_x \not\cong T'_{x'}$) he may need to swap graphs. We do not know if the upper bounds in Theorems 4.1 and 5.1 are valid if no alternations are allowed.

Acknowledgements

The authors thank the anonymous referee for helpful criticism.

References

[1] Barbour, A. D. (1982) Poisson convergence and random graphs. *Math. Proc. Camb. Phil. Soc.* **92** 349–359.

[2] Bollobás, B. (2001) *Random Graphs*, 2nd edn, Cambridge University Press.

[3] Chung, F. and Lu, L. (2001) The diameter of sparse random graphs. *Adv. Appl. Math.* **26** 257–279.

[4] Janson, S., Łuczak, T. and Ruciński, A. (2000) *Random Graphs*, Wiley–Interscience.

[5] Kim, J. H., Pikhurko, O., Spencer, J. and Verbitsky, O. (2005) How complex are random graphs in first order logic? *Random Struct. Alg.* **26** 119–145.

[6] Moon, J. W. (1968) On the maximum degree in a random tree. *Michigan Math. J.* **15** 429–432.

[7] Ore, O. (1962) *Theory of Graphs*, AMS, Providence, RI.

[8] Pikhurko, O., Spencer, J. and Verbitsky, O. (2005) Decomposable graphs and definitions with no quantifier alternation. To appear in *Europ. J. Combin.* The conference version (*EuroComb '05*) appeared in *Discrete Math. & Theoretical Comput. Sci.*, Vol. AE, pp. 25–30.

[9] Pikhurko, O., Spencer, J. and Verbitsky, O. (2006) Succinct definitions in the first order graph theory. *Annals Pure Appl. Logic* **139** 74–109.

[10] Pikhurko, O., Veith, H. and Verbitsky, O. (2006) The first order definability of graphs: Upper bounds for quantifier rank. *Discrete Appl. Math.* **154** 2511–2529.

[11] Pikhurko, O. and Verbitsky, O. (2005) Descriptive complexity of finite structures: Saving the quantifier rank. *J. Symb. Logic* **70** 419–450.

[12] Spencer, J. (2001) *The Strange Logic of Random Graphs*, Springer.

[13] Stanley, R. P. (1997) *Enumerative Combinatorics*, Cambridge University Press.

[14] Verbitsky, O. (2005) The first order definability of graphs with separators via the Ehrenfeucht game. *Theoret. Comp. Sci.* **343** 158–176.

https://doi.org/10.1017/S0963548306008376 Published online by Cambridge University Press