How Complex Are Random Graphs in First Order Logic?

Jeong Han Kim,[1] Oleg Pikhurko,[2] Joel H. Spencer,[3] Oleg Verbitsky[4]
[1] *Microsoft Research, One Microsoft Way, Redmond, Washington 98052;*
e-mail: jehkim@microsoft.com
[2] *Department of Mathematical Sciences, Carnegie Mellon University,*
Pittsburgh, Pennsylvania 15213-3890
[3] *Courant Institute, New York University, New York, New York 10012;*
e-mail: spencer@cs.nyu.edu
[4] *Department of Mechanics and Mathematics, Kyiv University, Ukraine;*
e-mail: oleg@ov.litech.net

Received 17 November 2003; accepted 31 August 2004
DOI 10.1002/rsa.20049
Published online 23 November 2004 in Wiley InterScience (www.interscience.wiley.com).

ABSTRACT: It is not hard to write a first order formula which is true for a given graph G but is false for any graph not isomorphic to G. The smallest number $D(G)$ of nested quantifiers in such a formula can serve as a measure for the "first order complexity" of G. Here, this parameter is studied for random graphs. We determine it asymptotically when the edge probability p is constant; in fact, $D(G)$ is of order $\log n$ then. For very sparse graphs its magnitude is $\Theta(n)$. On the other hand, for certain (carefully chosen) values of p the parameter $D(G)$ can drop down to the very slow growing function $\log^* n$, the inverse of the TOWER-function. The general picture, however, is still a mystery.
Random Struct. Alg., 26, 119–145, 2005

1. INTRODUCTION

In this paper we shall deal with sentences about graphs expressible in first-order logic. Namely, the vocabulary consists of the following symbols:

- *variables* (x, y, y_1, etc.);
- the *relations* $=$ (equality) and $\sim$ (the graph adjacency);

Correspondence to: O. Pikhurko; www.math.cmu.edu/~pikhurko/

- the *quantifiers* $\forall$ (universality) and $\exists$ (existence);
- the usual Boolean *connectives* ($\vee$, $\wedge$, $\neg$, $\Leftrightarrow$, and $\Rightarrow$).

These can be combined into first order *formulas* accordingly to the standard rules. A *sentence* is a formula without free variables. On the intuitive level it is perfectly clear what we mean when we say that a sentence *is true* on a graph G. This is denoted by $G \models A$; we write $G \not\models A$ for its negation (A *is false* on G). We do not formalize these notions; a more detailed discussion can be found in, e.g., [21, Section 1].

Please note that the variables represent vertices so the quantifiers apply to vertices only; i.e., we cannot express sentences like *"There is a set X having a given property."* (In fact, the language lacks any symbols to represent sets or functions.) We do not allow infinite sentences nor infinite graphs. As we do not go beyond first order logic, the standalone term "sentence" means a "first order sentence."

From a logician's point of view, the first order properties of graphs form a natural class of properties to study. For example, the so-called *zero-one laws* for random graphs have been extensively studied: see e.g., [7, 9, 11, 13–15, 19–21].

Of course, if $G \models A$ and $H \cong G$ (i.e. H is isomorphic to G), then $H \models A$. On the other hand, for any graph G it is possible to find a first order sentence A which *defines* G, that is, $G \models A$ while $H \not\models A$ for any $H \ncong G$. Indeed, let $V(G) = \{v_1, \dots, v_n\}$. The required sentence A could read:

> *"There are vertices $x_1, \dots, x_n$, all distinct, such that any vertex x_{n+1} is equal to one of these and $x_i \sim x_j$ iff $\{v_i, v_j\} \in E(G)$, $1 \le i < j \le n$."*

However, this sentence looks rather wasteful: we have $n + 1$ variables, the $\sim$-relation was used $\binom{n}{2}$ times, etc. Of a number of possible parameters measuring how complex A is, we choose here $D(A)$, the *quantifier depth* (or simply *depth*) which is the size of a longest sequence of embedded quantifiers. In the above example, $D(A) = n + 1$. This is a natural characteristic which appears, for example, in the analysis of algorithms for checking whether $G \models A$. Also, the depth function can be studied by using the so-called Ehrenfeucht game [8] (see Section 2 here). Following Pikhurko, Veith, and Verbitsky [17] (see also [18]), we let $D(G)$ be the smallest depth of a sentence defining G. It is a measure of how difficult it is to describe the graph G in first order logic.

A word of warning: the function $D(G)$ does not correlate very well with our everyday intuition of how complex the graph G is. Such are the limitations of the first order language that, for example, $D(K_n) = D(\overline{K}_n) = n + 1$ is the largest among all order-n graphs but what can be simpler than the complete or empty graph?!

The reason that $D(G)$ is large for $G = K_n$ is that this graph has a large *homogeneous set*, that is, a set $X \subset V$ such that any bijection $V \to V$ which is the identity on $V \setminus X$ is an automorphism of G. On the other hand, it is shown in [17] that if $D(G) > \frac{v(G)+5}{2}$, then G has a homogeneous set having at least $D(G) - 2$ vertices. However, the situation seems to get messy when we try to tackle graphs with $D(G) \ge \left(\frac{1}{2} - \varepsilon\right) v(G)$. Besides a large homogeneous set, there are other obstacles which may push the depth up: Cai, Fürer, and Immerman [6] constructed graphs G to define which we need $\Omega(v(G))$ nested quantifiers even if we add *counting* to first order logic. This is a rather drastic addition: for example, we need only two nested quantifiers to define K_n with counting, namely,

> *"There are precisely n vertices and every two of them are connected."*

The opposite approach was taken by Pikhurko, Spencer, and Verbitsky [16]: what is $q(n)$, the minimum $D(G)$ over all graphs G of order n? It turned out that $q(n)$ can be arbitrarily small in the following sense: for any recursive function $f : \mathbb{N} \to \mathbb{N}$ there is n such that $f(q(n)) < n$. If we try to "smooth" $q(n)$ by defining $q^*(n) = \max_{i \le n} q(i)$, then $q^*(n) = \Theta(\log^* n)$. Here, the *log-star* $\log^* n$ is the inverse to the TOWER-function, that is, the number of times we have to take the binary logarithm before we get below 1:

$$\log^* n = \min\{i \in \mathbb{N} : \log_2^{(i)} n < 1\}.$$

Such a behavior is surprising and intriguing. Having studied the two extreme cases, we concentrate now on what happens in a typical graph. More generally, we consider the standard Erdős-Rényi model $G \in \mathcal{G}(n,p)$, where p denotes the edge probability. Of course, we are interested in events occurring *whp* (with high probability, that is, with probability $1 - o(1)$ as $n \to \infty$). While a zero-one law studies the probability that a fixed sentence holds, we take a random graph and ask what the "simplest" sentence defining it is.

As $D(G) = D(\overline{G})$, we can assume without loss of generality that $p \le \frac{1}{2}$.

In Section 3 we study the case when $0 < p \le \frac{1}{2}$ is a constant and show that whp

$$D(G) = \log_{1/p} n + O(\ln \ln n). \qquad (1)$$

The case $p = \frac{1}{2}$ is always of particular interest: $G \in \mathcal{G}\big(n, \frac{1}{2}\big)$ is uniformly distributed among all graphs of order n. In Section 4, we have found a different line of argument (as far as the upper bound is concerned), which allowed us to pinpoint $D(G)$ down to at most 5 distinct values for infinitely many n. Unfortunately, this approach does not seem to work for $p \ne \frac{1}{2}$.

In Section 5 we show that for $p < \frac{1.19}{n}$, $D(G)$ is determined by the number of isolated vertices and therefore is of order $\Theta(n)$. Very recently, this result has been extended in [4] to all $p = O(n^{-1})$.

Rather surprisingly, for some carefully selected $p = p(n)$ the function $D(G)$ can be as small as $O(\log^* n)$. The reason is that the integer arithmetics can be modeled over the obtained random graphs while integers can be defined by first order sentences of very small depth. We do not present an exhaustive general theorem but give an example demonstrating this phenomenon when $p = n^{-1/4}$. On the other hand, the upper bound $O(\log^* n)$ is sharp, up to a multiplicative constant (cf. Theorem 20).

The first-order complexity of $G \in \mathcal{G}(n,p)$ for the general p remains a mystery. We refer the Reader to Section 7 for some concluding remarks and open questions.

2. THE EHRENFEUCHT GAME

For non-isomorphic graphs G and G' let $D(G, G')$ be the smallest quantifier depth of a first order sentence A *distinguishing* G from G' (that is, $G \models A$ while $G' \not\models A$). As the negation sign does not affect the depth, we have $D(G, G') = D(G', G)$.

Lemma 1. *For any graph G we have*

$$D(G) = \max\{D(G, G') : G' \not\cong G\}. \qquad (2)$$

Proof. Clearly, $D(G, G') \le v(G) + 1$, so the right-hand side of (2) is well-defined. Theorem 2.2.1 in [21] implies that all graphs can be split into finitely many classes so that any first-order sentence of depth at most $v(G)+1$ does not distinguish graphs within a class. For each class, except the one which contains G, pick a representative G' and let $A_{G'}$ be a minimum depth sentence distinguishing G from G'. The disjunction of these $A_{G'}$ proves the '$\le$'-inequality in (2).

The converse inequality is trivial. ∎

In the remainder of this section, we describe the Ehrenfeucht game which is a very useful combinatorial tool for studying $D(G,G')$. It was introduced by Ehrenfeucht [8]. Earlier, Fraïssé [10] suggested an essentially equivalent way to compute $D(G,G')$ in terms of partial isomorphisms between G and G'. A detailed discussion of the game can be found in [21, Section 2].

Let G and G' be two graphs. By replacing G' with an isomorphic graph, we can assume that $V(G)\cap V(G')=\emptyset$. The *Ehrenfeucht game* $\textsc{Ehr}_k(G,G')$ is played by two players, called *Spoiler* and *Duplicator* and consists of k rounds. For brevity, let us refer to Spoiler as "him" and to Duplicator as "her." In the ith round, $i=1,\dots,k$, 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. (A vertex may receive more than one mark.) At the end of the game (i.e., after k rounds) let $x_1,\dots,x_k$ be the vertices of G marked $1,\dots,k$ respectively, regardless of who put the label there; let $x'_1,\dots,x'_k$ 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 $\{x_i,x_j\}\in E(G)$ iff $\{x'_i,x'_j\}\in E(G')$ as well as that $x_i=x_j$ iff $x'_i=x'_j$. Otherwise, Spoiler wins.

The crucial relation is that for any non-isomorphic G and G' the smallest r such that Spoiler has a winning strategy in $\textsc{Ehr}_r(G,G')$ is equal to $D(G,G')$. In fact, an explicit winning strategy for Spoiler gives us an explicit sentence distinguishing G from G'.

If Spoiler can win the game, alternating between the graphs G and G' at most r times, then the corresponding sentence has the *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 depth of a sentence which defines G and has the alternation number at most r. Clearly, $D_r(G)=\max\{D_r(G,G') : G'\not\cong G\}$, 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, giving a qualitative strengthening of the obtained results. Therefore, we make the extra effort of computing the alternation number given by our strategies if the obtained r is really small.

Finally, let us make a few remarks on our terminology. When a player marks a vertex, we may also say that the player *selects* (or *claims*) the vertex. Duplicator *loses after i rounds* if the correspondence between $(x_1,\dots,x_i)$ and $(x'_1,\dots,x'_i)$ is not a partial isomorphism. (Of course, there is no point in continuing the game in this situation.)

3. CONSTANT EDGE PROBABILITY

As $D(G)=D(\overline{G})$, we can assume without loss of generality that $p\le\frac{1}{2}$. For brevity let us denote $q=1-p$. In this section we prove the following result.

Theorem 2. *Let p be a constant, $0<p\le\frac{1}{2}$. Let $G\in\mathcal{G}(n,p)$. Then whp*

$$-O(1)\le D(G)-\log_{1/p} n+2\log_{1/p}\ln n\le(2+o(1))\,\frac{\ln\ln n}{-p\ln p-q\ln q}. \tag{3}$$

The lower bound follows by observing that if for any disjoint $A, B \subset G$ with $|A| + |B| \le k$, there is a vertex y connected to everything in A but to nothing in B (this is called the *k-extension property* or the *k-Alice Restaurant property*), then $D(G) \ge k + 2$. The upper bound is obtained by some kind of recursion, where for every $x \in G$ we write a sentence A_x describing its neighborhood $\Gamma(x)$. Whp no two neighborhoods are isomorphic so A_x "defines" x and the final sentence A stipulates that the (unique) vertices satisfying A_x and A_y are connected if and only if $x, y \in G$ are. As each recursive step reduces the order by a factor of about $1/p$, the obtained sentence has depth around $\log_{1/p} n$. There are some technicalities to overcome. However, the combinatorial setting of the Ehrenfeucht games makes the proof more transparent and accessible.

Unfortunately, we have hardly any control on the alternation number in Theorem 2. The following result fills this gap by providing the defining sentences of a very restrictive form: no alternation at all. This, however, comes at the expense of increasing the depth by a constant factor.

Theorem 3. *Let p be a constant,* $0 < p < 1$*. Let* $G \in \mathcal{G}(n,p)$*. Then whp*

$$D_0(G) \le (2 + o(1)) \frac{\ln n}{-\ln(p^2 + q^2)}. \tag{4}$$

Remark. If we are happy to bound D_1 only, then the constant in (4) can be improved: In the proof (Section 3.3) we have to use Lemma 8 instead of Lemma 10.

3.1. The Lower Bound

To prove the lower bound in (3), we use the following lemma.

Lemma 4. *If G has the k-extension property, then* $D(G) \ge k + 2$.

Proof. Let $G' \not\cong G$ be another graph which has the k-extension property. (For example, we can take a random graph of large order.) Consider $\textsc{Ehr}_{k+1}(G, G')$. Duplicator's strategy is straightforward. If in the ith round Spoiler selects a previously marked vertex, Duplicator does the same in the other graph. Otherwise, she matches the adjacencies between x_i and $\{x_1, \dots, x_{i-1}\}$ to those between x'_i and $\{x'_1, \dots, x'_{i-1}\}$ by the k-extension property. ■

It is easy to show that whp $G \in \mathcal{G}(n,p)$, for constant $p \in \left(0, \frac{1}{2}\right]$, has the $\lfloor r \rfloor$-extension property with

$$r = \log_{1/p} n - 2\log_{1/p} \ln n + \log_{1/p} \ln(1/p) - o(1), \tag{5}$$

which gives us the required lower bound by Lemma 4. Indeed, for $k < r - \Theta(1)$, the expected number of 'bad' $A, B \subset V(G)$ with $|A| + |B| = k$ can be bounded by

$$\binom{n}{k} 2^k (1 - p^k)^{n-k} = \mathrm{e}^{k \ln n - p^k n + o(\ln^2 n)} = o(1).$$

3.2. The Upper Bound

Let $G = (V, E)$ be a graph. Let $\mathcal{V}_i$ consist of all ordered sequences of i pairwise distinct vertices of G. For $\mathbf{x} = (x_1, \dots, x_i) \in \mathcal{V}_i$ define

$$V_{\mathbf{x}} = \{y \in V \setminus \{x_1, \dots, x_i\} : \forall j \in [i]\ \{y, x_j\} \in E\},$$

and $G_{\mathbf{x}} = G[V_{\mathbf{x}}]$. We abbreviate $G_{x_1,\ldots,x_i} = G_{(x_1,\ldots,x_i)}$, etc. Let us agree that $\mathcal{V}_{-1} = \emptyset$, $\mathcal{V}_0 = \{()\}$ consists of the empty sequence, and $G_{()} = G$.

The following lemma specifies our global line of attack.

Lemma 5. *Suppose that a graph G, numbers $l \geq 0$ and $l_0 \geq 3$ satisfy all of the following conditions.*

1. *For any $\mathbf{x} \in \mathcal{V}_{l-1} \cup \mathcal{V}_l$ we have $D(G_{\mathbf{x}}) \leq l_0$.*
2. *For any $i \leq l-1$, $\mathbf{x} \in \mathcal{V}_i$, and distinct $y, z \in V_{\mathbf{x}}$, the following two conditions hold. Let $U = V_{\mathbf{x},y,z}$.*
   a. *Any injection $f : U \to V_{\mathbf{x},y}$ which embeds $G_{\mathbf{x},y,z}$ as an induced subgraph into $G_{\mathbf{x},y}$ is the identity mapping. (In particular, $G_{\mathbf{x},y,z}$ admits no non-trivial automorphism.)*
   b. *There is a vertex $v \in V_{\mathbf{x},y} \setminus U$ such that for any vertex $w \in V_{\mathbf{x},z} \setminus U$ we have $\Gamma(v) \cap U \neq \Gamma(w) \cap U$, where Γ denotes the neighborhood of a vertex.*
3. *For any $i \leq l-1$, $\mathbf{x} \in \mathcal{V}_i$, and distinct $y, z, w \in V_{\mathbf{x}}$, $G_{\mathbf{x},y,z}$ is not isomorphic to an induced subgraph of $G_{\mathbf{x},\mathbf{w}}$.*

Then $D(G) \leq l + l_0$.

Proof. Let us observe first that Condition 2 (or Condition 3) implies that

$$G_{x_1,\ldots,x_i,y} \not\cong G_{x_1,\ldots,x_i,z}, \quad \text{for any } i \leq l-1,\ (x_1,\ldots,x_i) \in \mathcal{V}_i, \text{ and distinct } y, z \in V_{x_1,\ldots,x_i}. \tag{6}$$

We prove the lemma by induction on l. If l is 0 or 1, then Condition 1 alone implies the claim. So, let $l \geq 2$. Let $G' = (V', E')$ be any graph which is not isomorphic to G.

Case 1: Suppose that there is $x \in V$ such that $G_x \not\cong G'_y$ for any $y \in V'$.

Spoiler selects this x. Let x' be the Duplicator's reply. The graph G_x satisfies all the assumptions of Lemma 5 with l decreased by 1. Spoiler will always play inside one of G_x or $G'_{x'}$. We can assume that Duplicator does the same for otherwise the adjacencies to x and x' do not correspond. As $G_x \not\cong G'_{x'}$, Spoiler can use the induction to win the $(G_x, G'_{x'})$-game in at most $l + l_0 - 1$ moves, as required. ∎

The same argument works if there is $x \in V'$ such that $G'_x \not\cong G_y$ for any $y \in V$.

Case 2: Suppose now that there are $x \in V$ and distinct $y', z' \in V'$ such that

$$G_x \cong G'_{y'} \cong G'_{z'}. \tag{7}$$

Spoiler selects $y' \in V'$. Assume that Duplicator replies with $y = x$, for otherwise $G_y \not\cong G'_{y'}$ by (6) and Spoiler proceeds as in Case 1. Now Spoiler selects z'; let $z \in V$ be the Duplicator's reply. We can assume that

$$G_{y,z} \cong G'_{y',z'}, \tag{8}$$

for otherwise Spoiler applies the inductive strategy to the $(G_{y,z}, G_{y',z'})$-game, where l is reduced by 2.

We show that Spoiler can win in at most 3 extra moves now. Let $U = V_{y,z}$ and $U' = V'_{y',z'}$. Spoiler selects the vertex $v \in V_y \setminus U$ given by Condition 2b. Let $v' \in V'_{y'} \setminus U'$ be the Duplicator's reply. By Condition 2a and (7)–(8) there is a bijection $g : V'_{y'} \to V'_{z'}$ which is the identity on U' and induces an isomorphism of $G'_{y'}$ onto $G'_{z'}$. Spoiler selects $w' = g(v')$. Whatever the reply $w \in G_z \setminus U$ of Duplicator is, $\Gamma(v) \cap U \neq \Gamma(w) \cap U$. But in G' we have $\Gamma(v') \cap U' = \Gamma(w') \cap U'$. Spoiler can point this difference with one more move into U. The total number of moves is $5 \le l + l_0$, as required. ∎

By (6), the only remaining case is the following.

Case 3: Suppose that there is a bijection $g : V \to V'$ such that for any $x \in V$ we have $G_x \cong G_{g(x)}$.

As $G \not\cong G'$, there are $y, z \in V$ such that g does not preserve the adjacency between y and z. Spoiler selects y. We can assume that Duplicator replies with $y' = g(y)$ for otherwise Spoiler proceeds as in Case 1. Now, Spoiler selects z to which Duplicator is forced to reply with $z' \neq g(z)$. Assume that $G_{y,z} \cong G_{y',z'}$ for otherwise Spoiler applies the inductive strategy for $l - 2$ to these graphs. But then $G_{y,z}$ is an induced subgraph of G_w, where $w = g^{-1}(z')$, contradicting Condition 3. ■

In order to finish the proof of Theorem 2, we apply Lemma 5 to a random graph $G \in \mathcal{G}(n,p)$.

Let $l = \log_{1/p} n - C \log_{1/p} \ln n - 1$ and $l_0 = C_0 \ln \ln n$, where C and C_0 are constants such that $C > 2$ and $C_0(-p \ln p - q \ln q) > C$. Let $m = np^{l+1} = \ln^C n$. Let $\varepsilon > 0$ be a small constant. Let n be sufficiently large.

Let $\mathcal{V} = \cup_{i=0}^{l+1} \mathcal{V}_i$. Observe that $|\mathcal{V}| \le \mathrm{e}^{(1+\varepsilon) \log_{1/p} n \ln n} = \mathrm{e}^{O(\ln^2 n)}$.

Lemma 6. *Whp for any $i \le l+1$ and $\mathbf{x} \in \mathcal{V}_i$ we have*

$$\big|\, |V_{\mathbf{x}}| - p^i n \,\big| \le \varepsilon p^i n. \tag{9}$$

Proof. Fix some $\mathbf{x} \in \mathcal{V}_i$. The size of $V_{\mathbf{x}}$ has the binomial distribution with parameters $(n-i, p^i)$. By Chernoff's bound ([1, Appendix A]), the probability p' that this $\mathbf{x}$ violates (9) is

$$p' \le 2\mathrm{e}^{-\varepsilon^2 np^i/3} \le 2\mathrm{e}^{-\varepsilon^2 m/3} = o(|\mathcal{V}|^{-1}). \tag{10}$$

Thus the expected number of "bad" $\mathbf{x}$'s is $o(1)$, giving the required. ■

Lemma 7. *Whp Condition 2 holds.*

Proof. Fix $i \le l - 1$, $\mathbf{x} \in \mathcal{V}_i$, and $y, z \in V_{\mathbf{x}}$. Let $U = V_{\mathbf{x},y,z}$, $W = V_{\mathbf{x},y}$, $u = |U|$, and $w = |W|$.

First we deal with Condition 2a. Take $j \in [1, u]$. Let g be any injection from U into W such that $|U_g| = j$, where $U_g = \{v \in U : g(v) \neq v\}$ consists of the elements moved by g. Let the same symbol g denote also the induced action on edges. Let E_g consist of those $e \in \binom{U}{2}$ such that $g(e) \neq e$. It is not hard to see that

$$|E_g| \ge \binom{j}{2} + j(u-j) - \frac{j}{2} = j\left(u - \frac{j}{2} - 1\right).$$

We can find a set $D \subset E_g$ of size at least $|E_g|/3$ such that $D \cap g(D) = \emptyset$. We do so greedily: choose any $e \in E_g$, move e to D, and remove $g(e), g^{-1}(e)$ from E_g (if they belong there). The probability that, for all $e \in D$, the 2-sets e and $g(e)$ are simultaneously edges or non-edges is $(p^2 + q^2)^{|D|}$ because these events are independent. This gives an upper bound on the probability that g induces an isomorphism.

Given j, there are at most $\binom{u}{j} w^j$ choices of g. The sequence $(\mathbf{x}, y)$, or $(\mathbf{x}, y, z)$, violates (9) with probability at most p', where p' is as in (10). Thus we can bound the probability that $\mathbf{x}, y, z$ violate Condition 2a by

$$2p' + \sum_{j=1}^{u} \binom{u}{j} w^j (p^2 + q^2)^{j(u-j/2-1)/3} < 2p' + (p^2 + q^2)^{\left(\frac{1}{3}-\varepsilon\right)m}.$$

Hence, the expected number of bad witnesses $\mathbf{x}, y, z$ is at most

$$|\mathcal{V}|\left(2p' + (p^2 + q^2)^{\left(\frac{1}{3}-\varepsilon\right)m}\right) = o(1),$$

giving the required by Markov's inequality.

To estimate the probability that Condition 2b fails, fix some $v \in W \setminus U$. The probability that some vertex of $V_{\mathbf{x},z} \setminus U$ has the same neighborhood in U as v is at most $(p^2 + q^2)^u$. We have $u > (1 - \varepsilon)m$ with probability at least $1 - p'$. Hence, v does not satisfy Condition 2b with probability at most

$$|V_{\mathbf{x},z}|\big(p' + (p^2 + q^2)^{(1-\varepsilon)m}\big) = o(|\mathcal{V}|),$$

finishing the proof. ∎

Condition 3 is verified similarly to the argument of Lemma 7. (The proof is, in a way, even easier because $|V_{\mathbf{x},y,z} \setminus V_w| = \Omega(m)$ whp.) All that remains to check is Condition 1. To deal with it, we need another strategic lemma. For a subset X of vertices of $G = (V, E)$, define the equivalence relation $\equiv_X$ on V, called the *X-similarity*, by $x \equiv_X y$ iff $x = y$ or $x, y \in V \setminus X$ satisfy $\Gamma(x) \cap X = \Gamma(y) \cap X$. This is an equivalence relation. Let

$$\mathcal{S}(X) = \{x \in V : \forall y \in V \; (y \neq x \;\Rightarrow\; y \not\equiv_X x)\} \supset X.$$

The vertices in $\mathcal{S}(X)$ are *sifted out* by X (that is, are uniquely determined by their adjacencies to X). We call X a *sieve* if $\mathcal{S}(X) = V$.

Lemma 8. *Let $X \subset V$. Define $Y = \mathcal{S}(X)$. If $\mathcal{S}(Y) = V$, then $D_1(G) \leq |X| + 3$.*

Proof. Let $G' \not\cong G$. First, Spoiler selects all of X. Let $X' \subset V'$ be the Duplicator's reply. Assume that Duplicator has not lost yet. For the notational simplicity let us identify X and X' so that $V \cap V' = X = X'$ and our both graphs coincide on X. Let $Y' = \mathcal{S}_{G'}(X)$.

It is not hard to see that Spoiler wins in at most two extra moves unless the following holds. For any $y \in Y \setminus X$ there is a $y' \in Y' \setminus X$ (and vice versa) such that $\Gamma(y) \cap X = \Gamma(y') \cap X$. Moreover, this bijective correspondence between Y and Y' induces an isomorphism between $G[Y]$ and $G'[Y']$.

Clearly, if Duplicator does not respect this correspondence, she loses immediately. Therefore, we may identify Y with Y'. Let $Z = V \setminus Y$ and $Z' = V' \setminus Y$. Let $z \in Z$ and define

$$W'_z = \{z' \in Z' : \Gamma(z') \cap Y = \Gamma(z) \cap Y\}.$$

If $W'_z = \emptyset$, Spoiler wins in at most two moves. First, he selects z. Let Duplicator reply with $z' \in Z'$. As the neighborhoods of z, z' in Y differ, Spoiler can highlight this by picking a vertex of Y. If $|W'_z| \geq 2$, then Spoiler selects some two vertices of W'_z and wins with at most one more move, as required.

Hence, we can assume that for any z we have $W'_z = \{f(z)\}$ for some $f(z) \in Z'$. It is easy to see that $f : Z \to Z'$ is in fact a bijection (otherwise Spoiler wins in two moves). As $G \not\cong G'$, the mapping f does not preserve the adjacency relation between some $y, z \in Z$. Now, Spoiler selects both y and z. Duplicator cannot respond with $f(y)$ and $f(z)$; by the definition of f Spoiler can win in one extra move. ∎

By Lemma 8, to complete the proof of Theorem 2 it suffices to verify that whp for any $\mathbf{x} \in \mathcal{V}_{l-1} \cup \mathcal{V}_l$ there is an $(l_0 - 3)$-set $X \subset V_{\mathbf{x}}$ such that, with respect to $H = G_{\mathbf{x}}$, $\mathcal{S}(Y) = U$, where $Y = \mathcal{S}(X)$ and $U = V_{\mathbf{x}}$. Let $k = l_0 - 3$ and $u = |U|$. Fix any $X \in \binom{U}{k}$. With probability at least $1 - p'$ we have $up^2 \leq (1 + \varepsilon)m$. Conditioned on this, $G_{\mathbf{x}}$ is still constructed by choosing its edges independently. The probability that a vertex $y \in U \backslash X$ belongs to Y is

$$\sum_{i=0}^{k} \binom{k}{i} p^i q^{k-i} \left(1 - p^i q^{k-i}\right)^{u-k-1}. \tag{11}$$

We want to bound this probability from below. Let, for example, $i_0 = pk - k^{1/2} \ln k$. For $i_0 \leq i \leq k$ we have $(1 - p^i q^{k-i})^u \geq 1 - \varepsilon$ by the definition of C_0. Chernoff's bound implies that $\sum_{i=i_0}^{k} \binom{k}{i} p^i q^{k-i} > 1 - \varepsilon$ as this sum corresponds to the Binomial distribution with parameters (k, p). Hence, the expression (11) is at least $(1 - \varepsilon)^2 > 1 - 2\varepsilon$ and the expectation

$$E[\,|Y|\,] > (1 - 2\varepsilon)u.$$

We construct the martingale $Y_0, \ldots, Y_{u-k}$, where we expose the vertices of $U \backslash X$ one by one and Y_i is the expectation of $|Y|$ after i vertices have been exposed. Changing edges incident to a vertex, we cannot decrease or increase $|Y|$ more than by two. By Azuma's inequality ([1, Theorem 7.2.1]), the probability that $|Y|$ drops, say, below $(1 - 3\varepsilon)u$ is at most $\mathrm{e}^{-\Omega(m)} = o(|\mathcal{V}|)$. Whp each Y has at least $(1 - 3\varepsilon)u$ elements. The following simple lemma completes our quest.

Lemma 9. *Let $\varepsilon = \varepsilon(p) > 0$ be sufficiently small. Whp for any $\mathbf{x} \in \mathcal{V}_{l-1} \cup \mathcal{V}_l$, every set $Y \subset V_{\mathbf{x}}$ of size at least $(1 - 3\varepsilon)u$, $u = |V_{\mathbf{x}}|$, is a sieve in $G_{\mathbf{x}}$.*

Proof. Let $\mathbf{x}$ satisfy (9). The expected number of bad triples (Y, y, z) (that is, the distinct vertices $y, z \in V_{\mathbf{x}} \backslash Y$ have the same neighborhood in a set Y of size at least $(1 - 3\varepsilon)u$) is

$$\sum_{i \leq 3\varepsilon u} \binom{u}{i} u^2 (p^2 + q^2)^{u-i} = o(|\mathcal{V}|).$$

The claim follows from (10). ∎

3.3. Games with No Alternations

Following our standard scheme, we first specify a graph property which ensures the desired bound on $D_0(G)$ and then show that a random graph satisfies this property whp.

Lemma 10. *Assume that in a graph $G = (V, E)$ we can find $X \subset V$ such that:*

1. *X is a sieve;*
2. *$G[X]$ has no nontrivial automorphism;*
3. *G has no other induced subgraph isomorphic to $G[X]$.*

Then $D_0(G) \le |X| + 2$.

Proof. Let G' be an arbitrary graph non-isomorphic to G. For some G' Spoiler plays all the time in G, for others he plays all the time in G'.

We first describe the strategy when Spoiler plays in G. Spoiler selects all vertices in X. Suppose that Duplicator managed to establish $\phi : X \to X'$, a partial isomorphism from G to G', where $X' \subset V'$ is the set of Duplicator's responses. Denote $Z = V(G)\setminus X$ and $Z' = V(G')\setminus X'$. We call two vertices, $v \in Z$ and $v' \in Z'$ *ϕ-similar* if the extension of ϕ which takes v to v' is a partial isomorphism from G to G'. Four cases are possible:

Case 1: The ϕ-similarity is a one-to-one correspondence between Z and Z'.

Case 2: There is $v \in Z$ without a ϕ-similar counterpart in Z'.

Case 3: There is $v' \in Z'$ without a ϕ-similar counterpart in Z.

Case 4: There are $v_1', v_2' \in Z'$ with the same ϕ-similar counterpart in Z.

In Case 1 there are $v_1, v_2 \in Z$ with adjacency different from the adjacency between their ϕ-similar counterparts in Z'. Spoiler selects v_1 and v_2 and wins. In Case 2 Spoiler wins by selecting the vertex v. In Cases 3 and 4 Spoiler fails in this way but plays differently from the very beginning.

Namely, if there exist X' and a partial isomorphism $\phi : X \to X'$ such that Cases 3 or 4 occur, Spoiler begins with selecting all vertices in X'. Duplicator is forced to reply in accordance with ϕ due to the conditions assumed for X. Then Spoiler selects the vertex v' in Case 3 or v_1' and v_2' in Case 4 and wins. ■

Lemma 11. *Let $\varepsilon > 0$ and $0 < p < 1$ be fixed. Let $G \in \mathcal{G}(n,p)$ and let $X \subset V$ be a fixed set of size $t \ge (2 + \varepsilon) \log_{1/r} n$, where $r = p^2 + q^2$. Then whp Conditions 1–3 of Lemma 10 hold.*

Proof. The expected number of vertices with the same neighborhood in X is at most $n^2 r^t = o(1)$, implying Condition 1. Conditions 2 and 3 follow from the following claim.

Claim 1. Whp no injective $g : X \to V$, with the exception of the identity mapping, preserves the adjacency relation.

Proof of Claim. Fix g. Let $k = |K|$ and $l = |L|$, where

$$K = \{x \in X : g(x) \notin X\},$$
$$L = \{x \in X : g(x) \in X\setminus\{x\}\},$$

As in the proof of Lemma 7 we can find a set $D \subset \binom{X\setminus K}{2}$ of size at least $l\,\frac{t-k-l/2-1}{3}$ such that $D \cap g(D) = \emptyset$. The latter property still holds if we enlarge D by the set of all elements of $\binom{X}{2}$ incident to K. Hence, the total probability of failure is at most

$$\sum_{0<k+l\le t} \binom{t}{k}\binom{t-k}{l} n^k t^l r^{l\,\frac{t-k-l/2-1}{3}+k(t-1)-\binom{k}{2}} = o(1),$$

completing the proof of the claim and the lemma. ∎

4. EDGE PROBABILITY 1/2

Here is the main result of this section.

Theorem 12. *Let $G \in \mathcal{G}\left(n, \frac{1}{2}\right)$. For infinitely many values of n we have whp*

$$D_2(G) \le \log_2 n - 2\log_2 \ln n + \log_2 \ln 2 + 6 + o(1). \tag{12}$$

Remark. The lower bound given by Lemma 4 and the case $p = \frac{1}{2}$ of (5) is by at most $5 + o(1)$ smaller than the upper bound in (12). This implies that $D(G)$ and $D_2(G)$ are concentrated on at most 6 different valued for such n. In Section 4.3 we will show that whp we have only 5 possible values.

Before we start proving Theorem 12, let us observe that for $p = \frac{1}{2}$ and an *arbitrary n* the upper bound (3) can be improved by using Lemma 8 to

$$D_1(G) \le \log_2 n - \log_2 \ln n + O(\ln\ln\ln n).$$

(Details are left to the interested reader.)

4.1. Spoiler's Strategy

Before we can specify the plan of our attack on Theorem 12, we have to give a few definitions.

Let $G = (V, E)$, $W \subset V$, and $u \in \mathbb{N}$. Building upon the notions defined before Lemma 8, let

$$\mathcal{S}_u(W) = \bigcup_{\substack{U \subset V\setminus W\\ |U|=u}} \left(\mathcal{S}(U \cup W)\setminus(U \cup W)\right).$$

In other words, a vertex $y \notin W$ belongs to $\mathcal{S}_u(W)$ iff there is a u-set $U \subset V\setminus(W \cup \{y\})$ such that $U \cup W$ sifts out y. Note that $\mathcal{S}(W) = \mathcal{S}_0(W) \cup W$.

Lemma 13. *Let $Y = \mathcal{S}_u(W)$. Suppose that $Y \cup W$ is a sieve in G and that no two vertices of Y have the same neighborhood in W. Then $D_2(G) \le u + w + 4$, where $w = |W|$.*

Proof. Let $G' = (V', E')$ be a graph non-isomorphic to G. We describe a strategy allowing Spoiler to win the game $\textsc{Ehr}_{u+w+4}(G, G')$.

Spoiler first claims W. Let Duplicator reply with $W' \subset V'$. Assume that she does not lose in this phase, establishing a partial isomorphism $f : W \to W'$. Recall that we call two vertices $v \in V$ and $v' \in V'$ *f-similar* if the extension of f taking v to v' is still a partial isomorphism from G to G'. Let $Y' = \mathcal{S}_u(W')$.

Claim 1. As soon as Spoiler moves inside $Y \cup Y'$ but Duplicator replies outside $Y \cup Y'$, Spoiler is able to win in the next $u + 1$ moves with 1 alternation between the graphs.

Proof of Claim. Assume for example that, while Spoiler selects $y \in Y$, Duplicator replies with $y' \notin Y'$. Spoiler selects some u-set U with y sifted out by $U \cup W$. Let the reply to it be U'. By the assumption on y', there is another vertex z' with the same adjacencies to $U' \cup W'$. Spoiler selects z' and wins. ▎

Claim 2. If Y' contains two vertices with the same adjacencies to W', then Spoiler is able to win the game in $w + u + 3$ moves with at most 2 alternations.

Proof of Claim. Assume that y' and z' are both in Y' and have the same adjacencies to W'. Spoiler selects these two vertices. In order not to lose immediately, Duplicator is forced to reply at least once outside Y. Spoiler wins in the next $u + 1$ moves according to Claim 1. ▎

Assume therefore that all vertices in Y' have pairwise distinct neighborhoods in W'. This assumption and Claim 1 imply that either the f-similarity determines a one-to-one correspondence between Y and Y' or Spoiler is able to win the game in $w + u + 3$ moves with at most 2 alternations. We will assume the first alternative. Extend f to a map from $W \cup Y$ onto $W' \cup Y'$ accordingly to the f-similarity correspondence between Y and Y'.

Claim 3. Suppose that Duplicator failed to respect the bijection f after a Spoiler's move into $Y \cup Y'$. Then Spoiler can win in at most $u + 1$ extra moves, during which he alternates at most once.

Proof of Claim. Suppose that the previous move x of Spoiler was in G, for example. Clearly, the Duplicator's response x' cannot belong to Y' because $f(x)$ is the only vertex in Y' with the required W'-adjacencies. Spoiler applies the strategy of Claim 1. ▎

Claim 4. If $f : W \cup Y \to W' \cup Y'$ is not a partial isomorphism from G to G', then Spoiler is able to win the game in $w + u + 3$ moves with 1 alternation.

Proof of Claim. Assume, for example, that $\{y_1, y_2\} \in \binom{Y}{2}$ is an edge while $\{f(y_1), f(y_2)\}$ is not. Spoiler picks y_1 and y_2. Duplicator cannot reply with $f(y_1)$ and $f(y_2)$ so Spoiler wins in at most $w + 2 + (u + 1) = w + u + 3$ moves by Claim 3. ▎

Assume therefore that $f : W \cup Y \to W' \cup Y'$ is a partial isomorphism. Denote $R = V \setminus (W \cup Y)$ and $R' = V' \setminus (W' \cup Y')$.

Claim 5. As soon as Spoiler moves inside $R \cup R'$ but Duplicator fails to reply with an f-similar vertex in $R \cup R'$ (in the other graph), Spoiler can win in at most $u + 2$ extra moves, during which he alternates at most once.

Proof of Claim. If Duplicator replies with a vertex $x \in Y \cup Y'$, in the next move Spoiler marks the f-mate of x and then applies the strategy of Claim 3. If she replies in $R \cup R'$ but not with an f-similar vertex, Spoiler highlights this in one more move and again uses Claim 3. ▎

Claim 6. If $W' \cup Y'$ is not a sieve in G', then Spoiler is able to win the game in $w + u + 4$ moves with 2 alternations.

Proof of Claim. Spoiler picks two witnesses $z'_1, z'_2 \in R'$ with the same adjacencies to $W' \cup Y'$. If at least one of the corresponding replies z_1, z_2 is not in R, Spoiler applies the strategy of Claim 5. Otherwise z_1 and z_2 belong to different $W \cup Y$-similarity classes and there is a vertex $x \in W \cup Y$ adjacent to exactly one of z_1 and z_2. If $x \in W$, Spoilers wins immediately. If $x \in Y$, Spoiler picks $f(x) \in Y'$. Duplicator cannot respond with x. Now Spoiler wins the game in at most $u + 1$ extra moves by Claim 3. ▮

In the rest of the proof we suppose that $W' \cup Y'$ is a sieve. This assumption and Claim 5 imply that either the f-similarity determines a one-to-one correspondence between R and R' or Spoiler is able to win the game in $w + u + 3$ moves with 2 alternations. Let us assume the first alternative. Extend f to the whole of V accordingly to the f-similarity correspondence between R and R'. Thus f is a bijection between V and V' now. As G and G' are not isomorphic, f does not preserve the adjacency for some $\{y_1, y_2\} \in \binom{V}{2}$. Spoiler selects y_1 and y_2. If Duplicator replies with $f(y_1)$ and $f(y_2)$, she loses immediately. Otherwise, Spoiler applies the strategy of Claim 5 and wins, having made totally at most $u + w + 4$ moves and 1 alternation. ▪

4.2. The Probabilistic Part

Let k be given. For simplicity let us assume that k is even. Define

$$f(n,k) = \binom{n - \frac{k}{2}}{\frac{k}{2}}(n-k)(1-2^{-k})^{n-k-1}.$$

Basic asymptotics show that for $n = \Theta(k^2\, 2^k)$ we have $\frac{f(n+1,k)}{f(n,k)} \approx 1$, and thus we can find $n = \left(\frac{\ln 2}{2} + o(1)\right) k^2 2^k$ such that

$$f(n,k) = (10 + o(1))\ \log_2 n.$$

We fix this n. Routine calculations show that

$$k \le \log_2 n - 2\log_2 \ln n + \log_2 \ln 2 + 1 + o(1). \tag{13}$$

Let A be a fixed $\frac{k}{2}$-subset of $G \in \mathcal{G}(n, \frac{1}{2})$. Let $\mathcal{U}$ consist of pairs (U, y), where U is a k-set containing A and $y \in V \backslash U$. For $(U, y) \in \mathcal{U}$, let $I(U, y)$ denote the indicator random variable for the event $y \in \mathcal{S}(U)$. We define

$$X = \sum_{(U,y)\in\mathcal{U}} I(U,y),$$

$$M = |\mathcal{U}| = \binom{n - k/2}{k/2}(n-k),$$

$$p = E[I(U,y)\,] = (1-2^{-k})^{n-k-1},$$

We further set

$$\mu = E[X] = Mp = f(n,k) = (10 + o(1))\log_2 n. \tag{14}$$

The idea behind these definitions is that we try to apply Lemma 13 for $W = A$ and $u = \frac{k}{2}$. Then μ is the expected number of ways to construct a vertex of $\mathcal{S}_u(W)$. Our proof works only if μ is neither too big nor too small, that is, for some special values of n only. We do not know if $D(G)$ can pinned down to $O(1)$ distinct values for an arbitrary n.

As $k \approx \log_2 n$ we have

$$M \approx \frac{n^{\frac{k}{2}+1}}{(k/2)!} = n^{\frac{k}{2}(1+o(1))}.$$

As $\mu = n^{o(1)}$ we further have

$$p \approx e^{-n2^{-k}} = n^{-\frac{k}{2}(1+o(1))}. \tag{15}$$

Lemma 14. *For distinct* $(U_1, y_1), (U_2, y_2) \in \mathcal{U}$,

$$E[I(U_1, y_1)I(U_2, y_2)] \leq n^{-\frac{3k}{4}(1+o(1))}. \tag{16}$$

When $|U_1 \cap U_2| < \frac{9k}{10}$,

$$E[I(U_1, y_1)I(U_2, y_2)] \leq E[I(U_1, y_1)]E[I(U_2, y_2)] \left(1 + O\left(n^{-\frac{k}{10}(1+o(1))}\right)\right). \tag{17}$$

Proof. Condition on the adjacency patterns of y_1, y_2 to U_1, U_2, respectively. Let z be any vertex not in $U_1 \cup U_2 \cup \{y_1, y_2\}$. Suppose (the main case) $U_1 \neq U_2$. Then

$$\Pr[(z \equiv_{U_1} y_1) \wedge (z \equiv_{U_2} y_2)] \leq 2^{-k-1} \tag{18}$$

as the adjacency pattern of z to $U_1 \cup U_2$ is then determined. When $U_1 = U_2$ the adjacency patterns of y_1, y_2 to U_1 must be different as otherwise $I(U_1, y_1) = I(U_2, y_2) = 0$. Then it would be impossible to have $z \equiv_{U_1} y_1$ and $z \equiv_{U_2} y_2$ so (18) still holds. By inclusion-exclusion

$$\Pr[(z \equiv_{U_1} y_1) \vee (z \equiv_{U_2} y_2)] \geq 2 \cdot 2^{-k} - 2^{-k-1} = 3 \cdot 2^{-k-1}.$$

If $I(U_1, y_1)I(U_2, y_2) = 1$, then this fails for all such z. But these events are mutually independent. Thus, by (15)

$$E[I(U_1, y_1)\, I(U_2, y_2)] \leq (1 - 3 \cdot 2^{-k-1})^{n-2k-2} = n^{-\frac{3k}{4}(1+o(1))}, \tag{19}$$

giving the required.

Suppose further that $|U_1 \cap U_2| < \frac{9k}{10}$. Again let z be any vertex not in $U_1 \cup U_2 \cup \{y_1, y_2\}$ and condition on the adjacency patterns of y_1, y_2 to U_1, U_2 respectively. Now

$$\Pr[(z \equiv_{U_1} y_1) \wedge (z \equiv_{U_2} y_2)] \leq 2^{-\frac{11k}{10}}$$

as this event requires z to have a given adjacency pattern to $U_1 \cup U_2$. Again by inclusion-exclusion

$$\Pr[(z \equiv_{U_1} y_1) \vee (z \equiv_{U_2} y_2)] \geq 2 \cdot 2^{-k} - 2^{-\frac{11k}{10}}.$$

As with (19) we deduce

$$E[I(U_1, y_1)I(U_2, y_2)] \leq \left(1 - 2 \cdot 2^{-k} + 2^{-\frac{11k}{10}}\right)^{n-2k-2}.$$

Now we want to compare this to $E[I(U_1,y_1)]E[I(U_2,y_2)]$. We have

$$\left(1-2\cdot 2^{-k}+2^{-\frac{11k}{10}}\right)^{n-2k-2}=((1-2^{-k})^{n-k-1})^2\left(1+O\left(n2^{-\frac{11k}{10}}\right)\right),$$

yielding (17). ∎

Lemma 15.

$$\mathrm{Var}[X]=O(E[X]). \tag{20}$$

Proof. As $X=\sum I(U,y)$, the sum of indicator random variables, we employ the general bound

$$\mathrm{Var}[X]\le E[X]+\sum \mathrm{Cov}[I(U_1,y_1),I(U_2,y_2)], \tag{21}$$

the sum over distinct $(U_1,y_1),(U_2,y_2)\in\mathcal{U}$. The first term is μ. Consider the sum of the covariances satisfying $|U_1\cap U_2|>\frac{9k}{10}$. There are M choices for (U_1,y_1). For a given U_1 there are $n^{\frac{k}{10}(1+o(1))}$ choices for (U_2,y_2) and

$$\sum E[I(U_1,y_1)I(U_2,y_2)]\le Mn^{\frac{k}{10}(1+o(1))}n^{-\frac{3k}{4}(1+o(1))}=o(1). \tag{22}$$

As the covariance of indicator random variables is at most the expectation of the product,

$$\sum \mathrm{Cov}[I(U_1,y_1),I(U_2,y_2)]=o(1), \tag{23}$$

where in (22)–(23) the sum is restricted to $|U_1\cap U_2|>\frac{9k}{10}$. When $|U_1\cap U_2|\le\frac{9k}{10}$, we have from (17) that

$$\mathrm{Cov}[I(U_1,y_1),I(U_2,y_2)]=O\Big(E[I(U_1,y_1)]E[I(U_2,y_2)]n^{-\frac{k}{10}(1+o(1))}\Big).$$

Hence $\big($the sum over $|U_1\cap U_2|\le\frac{9k}{10}\big)$

$$\sum \mathrm{Cov}[I(U_1,y_1),I(U_2,y_2)]=O\Big(n^{-\frac{k}{10}(1+o(1))}\Big)\sum E[I(U_1,y_1)]\,E[I(U_2,y_2)]$$

But $\sum E[I(U_1,y_1)]E[I(U_2,y_2)]$ over *all* $(U_1,y_1),(U_2,y_2)\in\mathcal{U}$ is precisely $\mu^2=O(\ln^2 n)$. This becomes absorbed in the $n^{-\frac{k}{10}}$ term and $\big($the sum again over $|U_1\cap U_2|\le\frac{9k}{10}\big)$

$$\sum \mathrm{Cov}[I(U_1,y_1),I(U_2,y_2)]=O\left(n^{-\frac{k}{10}(1+o(1))}\right). \tag{24}$$

In particular, all covariances in (21) together add up to $o(1)$ and so we actually have the stronger result $\mathrm{Var}[X]\le E[X]+o(1)$. ∎

Lemma 16. *Whp every pair $(U_1,y_1)\neq(U_2,y_2)$ from $\mathcal{U}$ with $I(U_1,y_1)=I(U_2,y_2)=1$ satisfies*

1. $U_1\cap U_2=A$;
2. $y_2\notin U_1$;
3. $y_1\neq y_2$;

4. *$y_1 \not\equiv_A y_2$;*
5. *For $u_1, u_2 \in U_1$ with $u_1 \neq u_2$, $u_1 \not\equiv_A u_2$;*
6. *For $u_1 \in U_1, u_2 \in U_2$, $u_1 \not\equiv_A u_2$.*

Proof. From (22) the expected number of pairs $(U_1, y_1) \neq (U_2, y_2)$ with $I(U_1, y_1) = I(U_2, y_2) = 1$ and $|U_1 \cap U_2| > \frac{9k}{10}$ is $o(1)$. The total number of pairs $(U_1, y_1) \neq (U_2, y_2)$ with $(U_1 \cap U_2) \backslash A \neq \emptyset$ is less than $M^2 \frac{k^2}{n}$. For each with $|U_1 \cap U_2| \leq \frac{9k}{10}$ a weak form of (17) gives that $E[I(U_1, y_1)\ I(U_2, y_2)] \leq 2p^2$. Hence the expected number of such pairs with $I(U_1, y_1) = I(U_2, y_2) = 1$ is bounded from above by $M^2 \frac{k^2}{n}(2p^2) = O((\ln^4 n)/n) = o(1)$. Hence the probability that Property 1 fails is $o(1)$.

For Property 2 we first employ Property 1 and restrict attention to $U_1 \cap U_2 = A$. The number of $(U_1, y_1), (U_2, y_2)$ with $y_2 \in U_1$ is less than $M^2 \frac{k}{n-k}$ and for each $E[I(U_1, y_1)\ I(U_2, y_2)] \approx p^2$ so the expected number with $I(U_1, y_1) I(U_2, y_2) = 1$ is less than $\approx (Mp)^2 \frac{k}{n}$ which is $O((\ln^3 n)/n) = o(1)$.

For Property 3 we first employ Property 1 and restrict attention to $U_1 \cap U_2 = A$. The number of such $(U_1, y_1), (U_2, y_2)$ with $y_1 = y_2$ is about $M^2 n^{-1}$ and for each such $\Pr[I(U_1, y_1) = I(U_2, y_2) = 1] \approx p^2$ so the expected number of violations of Property 3 is around $(Mp)^2 n^{-1} = \mu^2 n^{-1} = o(1)$.

For Property 4 we first employ Properties 1–3 and restrict attention to pairs satisfying those conditions. The number of such pairs is $\approx M^2$. For each such $\Pr[I(U_1, y_1) = I(U_2, y_2) = 1] \approx p^2$ and, conditioned on this, $\Pr[y_1 \equiv_A y_2] \approx 2^{-k/2}$. Hence the expected number of pairs violating Property 4 is $\approx (Mp)^2 2^{-k/2} = \mu^2 2^{-k/2}$, which is again $o(1)$.

For Property 5 there are M choices of U_1, y_1 and $O(k^2)$ choices for u_1, u_2. For each $\Pr[I(U_1, y_1) = 1] = p$, $\Pr[u_1 \equiv_A u_2] = 2^{-k/2}$ and these events are independent so the expected number of violations of Property 5 is $O(Mpk^2 2^{-k/2})$, which is $o(1)$.

For Property 6 we restrict attention to those cases satisfying Properties 1–4. There are less than M^2 choices of U_1, y_1, U_2, y_2 and $O(k^2)$ choices for u_1, u_2. For each Choice $\Pr[I(U_1, y_1) = I(U_2, y_2) = 1] \approx p^2$, $\Pr[u_1 \equiv_A u_2] = 2^{-k/2}$ and these events are independent so the expected number of violations of Property 6 is $O(M^2 p^2 k^2 2^{-k/2})$, which is $o(1)$. ■

Lemma 17. *Let $G \in \mathcal{G}(n, \frac{1}{2})$ and A be a fixed subset of $\frac{k}{2}$ vertices, as above. Let Z denote the union of all sets $U - A$ where some $I(U, y) = 1$ and let $S = \mathcal{S}_{k/2}(A)$ denote the set of all such vertices y. Let $R = V \backslash (A \cup Z \cup S)$. Then whp:*

1. *All $y_1, y_2 \in S$ have $y_1 \not\equiv_A y_2$.*
2. *All $u_1, u_2 \in Z$ have $u_1 \not\equiv_A u_2$.*
3. *There are no distinct $z_1, z_2, z_3, z_4 \in R$ with $z_1 \equiv_S z_2$ and $z_3 \equiv_S z_4$.*
4. *There are no distinct $z_1, z_2, z_3 \in R$ with $z_1 \equiv_S z_2 \equiv_S z_3$.*

Proof. The first two statements are Conclusions 4–6 of Lemma 16. We concentrate on showing Property 3 as Property 4 is similar. Set

$$l = \mu - \mu^{0.6}. \tag{25}$$

Let Y denote the number of l-sets $\{(U_i, y_i) : 1 \leq i \leq l\}$ (counting permutations of the (U_i, y_i) as the same) and z_1, z_2, z_3, z_4 satisfying:

- $I(U_i, y_i) = 1$ for $1 \leq i \leq l$.
- The $U_i - A$ are disjoint, the y_i are distinct, and no $y_i \in U_j$.

- $z_1, z_2, z_3, z_4 \in R$, where R denotes all vertices except the U_i and the y_i.
- $z_1 \equiv_S z_2$ and $z_3 \equiv_S z_4$, where we set $S = \{y_1, \ldots, y_l\}$.

We bound $E[Y]$. There are less than $M^l/l!$ choices for the (U_i, y_i) and n^4 choices for the z_1, z_2, z_3, z_4. Fix those choices. Set $R^- = R \backslash \{z_1, z_2, z_3, z_4\}$, and let $z \in R^-$. For each $1 \le i \le l$

$$\Pr[z \equiv_{U_i} y_i] = 2^{-k}.$$

For each $1 \le i, j \le l$, as $|U_i \cap U_j| = \frac{k}{2}$,

$$\Pr[(z \equiv_{U_i} y_i) \wedge (z \equiv_{U_j} y_j)] \le 2^{-\frac{3k}{2}}.$$

We apply the Bonferroni inequality, in the form that the probability of a disjunction is at least the sum of the probabilities minus the sum of the pairwise probabilities:

$$\Pr\left[\vee_{i=1}^{l} z \equiv_{U_i} y_i\right] \ge l2^{-k} - \binom{l}{2} 2^{-\frac{3k}{2}}.$$

These events are independent over the $z \in R^-$ as they involve different adjacencies. Let OK denote the event that no $z \equiv_{U_i} y_i$ for any $z \in R^-$ and $1 \le i \le l$. The independence gives:

$$\Pr[\mathrm{OK}] \le \left(1 - l\,2^{-k} + \binom{l}{2} 2^{-\frac{3k}{2}}\right)^{n - l(1+\frac{k}{2}) - \frac{k}{2}}.$$

We bound

$$1 - l\,2^{-k} + \binom{l}{2} 2^{-\frac{3k}{2}} \le (1 - 2^{-k})^l \,(1 + n^{-1.1})$$

and

$$n - l\left(1 + \frac{k}{2}\right) - \frac{k}{2} \le n - k - 1,$$

so that

$$\Pr[\mathrm{OK}] \le p^l (1 + n^{-1.1})^n \le p^l (1 + o(1)).$$

Our saving comes from

$$\Pr[z_1 \equiv_S z_2] = \Pr[z_3 \equiv_S z_4] = 2^{-l}.$$

The adjacencies on the z_i to S are independent of the event OK. But

$$\wedge_{i=1}^{l} I(U_i, y_i) = 1 \Rightarrow \mathrm{OK}.$$

Thus

$$\Pr\left[\left(\wedge_{i=1}^{l} I(U_i, y_i)\right) \wedge (z_1 \equiv_S z_2) \wedge (z_3 \equiv_S z_4)\right] \le p^l\, 2^{-2l} (1 + o(1)).$$

Putting this together,

$$E[Y] \le \frac{M^l}{l!} n^4 p^l 2^{-2l} (1 + o(1)).$$

Recall that $Mp = \mu \approx 10 \log_2 n$. The function $\mu^x/x!$ hits a maximum at $x = \mu$ where it is less than e^{μ}. Thus

$$\frac{(Mp)^l}{l!} \le \mathrm{e}^{\mu}.$$

Hence

$$E[Y] \le e^{\mu} n^4 2^{-2l}.$$

We have selected $l \approx \mu$ so that

$$e^{\mu} 2^{-2l} = \left(\frac{e}{4}\right)^{\mu(1+o(1))} = n^{-K(1+o(1))},$$

where $K = -10\log_2(e/4) > 4$. We deduce

$$E[Y] = o(1),$$

so that almost surely there is no such l-tuple. Recall that X was the total number of (U, y) with $I(U, y) = 1$. As $E[X] = \mu$ and, from (20), $\mathrm{Var}[X] = O(\mu)$ with probability $1 - o(1)$ we have $X \ge l$. Further, Lemmas 14 and 16 give that whp the extensions have Properties 1–2. Thus whp there exists a family of (U_i, y_i) of size l which satisfies Properties 1–2. But also whp any such family of size l will satisfy Properties 3 and 4. So whp there is such a family. The expansion of the family to all (U, y) with $I(U, y) = 1$ retains Properties 3–4 as the set S is just getting larger. So the theorem is proved. ▪

4.3. Putting All Together

We can now finish the proof of Theorem 12. By Lemma 17 we have whp that all $A \cup \mathcal{S}_{k/2}(A)$-similarity classes are singletons except possibly one 2-element class $\{x, y\}$. If we let $W = A \cup \{x\}$ and $u = \frac{k}{2}$, then clearly G satisfies all the assumptions of Lemma 13, which implies that $D_2(G) \le k + 5$, giving the required by (13).

Finally, let us justify the Remark after Theorem 12. Recall that given k we have chosen n so that $f(n, k) \approx 10\log_2 n$ and deduced that $D_2(G) \le k + 5$ whp. The probability that the $(k - 1)$-extension property fails for G is at most

$$\binom{n}{k-1}(1 - 2^{-k+1})^{n-k} = e^{k\ln n - k\ln k + k - 2^{-k+1}n + o(k)} = f^2(n, k)\, 2^{-k+o(k)}\, n^{-2} = o(1).$$

By Lemma 4, $k + 1 \le D(G)$. Thus, $D(G)$ and $D_2(G)$ are concentrated on at most 5 different values.

5. SPARSE RANDOM GRAPHS

The following lemma helps us to deal with very sparse random graphs. Let $t_k = t_k(G)$ be the number of components of G which are order-k trees. (Thus $t_1(G)$ is the number of isolated vertices.) For a graph F, let $c_F(G)$ be the number of components isomorphic to F.

Lemma 18. *Suppose that for any connected component F of a graph G we have*

$$c_F(G) + v(F) \le t_1(G) + 1. \tag{26}$$

Then $D(G) = D_1(G) = t_1(G) + 2$ unless G is an empty graph (when $D(G) = D_0(G) = v(G) + 1$).

Proof. Assume that $e(G) \ne 0$.

The lower bound on $D(G)$ follows by considering G' which is obtained from G by adding an isolated vertex. The graphs G and G' are isomorphic as far as non-isolated vertices are concerned. The best strategy for Spoiler is to pick $t_1(G)+1$ isolated vertices in G' and, by making one more move in G, to show that at least one of the Duplicator's responses is not an isolated vertex.

On the other hand, let $G' \not\cong G$. There must be a connected graph F such that $c_F(G) \neq c_F(G')$, say $c_F(G) < c_F(G')$. Spoiler picks one vertex from some $c_F(G)+1$ F-components of G'. If a move x of Duplicator falls into the same component of G as some her previous move y, then Spoiler switches to G and begins claiming a contiguous path from x to y; he wins in at most $v(F)$ moves by either connecting x to y or by claiming a path of length $v(F)+1$.

Otherwise, Duplicator must have selected a vertex inside a component C of G which is not isomorphic to F. As soon as this happens, Spoiler wins by growing a connected set inside the larger component of the two, in at most $v(F)$ moves.

The total number of moves does not exceed $(c_F(G)+1)+v(F) \leq t_1(G)+2$ (while we have only one alternation), as required. ■

Theorem 19. *Let $\varepsilon > 0$ be fixed. Let $p = p(n) \leq (\alpha - \varepsilon)n^{-1}$, where $\alpha = 1.1918\ldots$ is the (unique) positive root of the equation*

$$s\mathrm{e}^{-s} = \alpha\mathrm{e}^{-\alpha}, \quad \textit{where we denote } s := \alpha - \alpha\mathrm{e}^{-\alpha}.$$

Then whp $G \in \mathcal{G}(n,p)$ satisfies the condition (26). *In this range, whp*

$$D(G) = D_1(G) = (\mathrm{e}^{-pn} + o(1))\, n. \tag{27}$$

Proof. It is easy to compute the expectation of $t_k(G)$ for $G \in \mathcal{G}(n,p)$:

$$\lambda_k = E[t_k] = \binom{n}{k} k^{k-2} p^{k-1} q^{k(n-k)+\binom{k}{2}-k+1}.$$

Let $c = pn$. For a fixed k we have $\lambda_k = (f_k + o(1))n$, where $f_k := \frac{c^{k-1}k^{k-2}}{k!\,\mathrm{e}^{ck}}$. We have

$$\frac{f_{k+1}}{f_k} = c\mathrm{e}^{-c} \times \left(1 + \frac{1}{k}\right)^{k-2}.$$

The first factor $c\mathrm{e}^{-c}$ is at most $1/\mathrm{e}$ (maximized for $c = 1$). Unexciting algebraic calculations show that the second factor is monotone increasing for $k \geq 1$ and approaches e in the limit. This implies that the sequence f_k is decreasing in k. (In particular, f_1 is strictly bigger than any other f_i, $i \geq 2$.)

Also, $f_1 = \mathrm{e}^{-c} > 0.3$ for $0 \leq c \leq \alpha$.

Theorems 5.7 and 6.11 in Bollobás [5] describe the structure of a typical G for $p = O(n^{-1})$. In particular, it implies that there is a constant K such that whp at least $0.9n$ vertices of G belong either to tree components of orders at most K or to the giant component. The giant component (for $c > 1$) has order $(1 - \frac{s}{c} + o(1))\, n$, where s is the only solution of $s\mathrm{e}^{-s} = c\mathrm{e}^{-c}$ in the range $0 < s \leq 1$. It is routine to see that $f_1 > 1 - \frac{s}{c}$. (In fact, $c = \alpha$ is the root of $f_1 = 1 - \frac{s}{c}$; this is where α comes from.)

A theorem of Barbour [3] (Theorem 5.6 in [5]) implies that, for any $k \leq K$, we have whp

$$|t_k(G) - \lambda_k| \leq o(n). \tag{28}$$

Now, we have all the ingredients we need to check (26). Let $F \subset G$ be any connected component. If F is the giant component of G, then $c_F(G) = 1$, but, as we have seen, $v(F) \leq t_1(G)$, so (26) holds. So we can assume that $v(F) = o(n)$. If $v(F) > K$, then $c_F(G) + v(F) \leq \frac{0.1n}{K} + K < t_1(G)$. If F is a tree with $k \in [2, K]$ vertices, then (28) and the inequality $f_1 > f_k$ imply the required. Finally, it remains to assume that the component F of order at most K contains a cycle. But the expected number of such components is at most $\sum_{i=2}^{K} \binom{n}{i} p^i i^{i-1} \binom{i}{2} = O(1)$. Markov's inequality implies that whp no such F violates (26).

■

Of course, the value of $t_1(G)$ can be estimated more precisely for some p than we did in Theorem 19. Without going into much details, let us describe some of the cases here. Let ω be any function of n which (arbitrarily slowly) tends to the infinity with n.

If $n^2p \to 0$, then whp we have isolated vertices and edges only. The distribution of $t_2(G) = e(G)$ approaches the Poisson distribution $\mathcal{P}_{\lambda_2}$. Hence, we have whp that $n - \omega < D(G) \leq n + 1$.

Suppose that $n^2p \not\to 0$ but $pn \to 0$. The expected number of vertices in components of order at least 3 is at most $n\binom{n}{2}3p^2 = o(\lambda_2)$. By Markov's inequality, whp we have $o(\lambda_2)$ such vertices. On the other hand, the distribution of $t_2(G)$ is $o(1)$-close to $\mathcal{P}_{\lambda_2}$ (Theorem 5.1 in [5]). Hence,

$$\left| D(G) - n + \frac{\lambda_2}{2} \right| \leq o(\lambda_2) + \omega.$$

Observe that there is no phase transition in the behavior of $D(G)$ at $p \approx \frac{1}{n}$. This should not be surprising: $D(G)$ is determined by $t_1(G)$ in this range. The more recent results in [4] imply that (27) holds for any $p = O(n^{-1})$.

6. MODELING ARITHMETICS ON GRAPHS

In this section we consider $D(G)$ for the random graph $G \in \mathcal{G}(n,p)$ where $p = n^{-1/4}$. We expect that our results would hold for $p = n^{-\alpha}$ for any *rational* $\alpha \in (0, 1)$, but this would require considerable technical work so we are content with this one case. In [21, Section 8] it was shown, for $\alpha = \frac{1}{3}$, that there was an arithmetization of certain sets that led to non-convergence and non-separability results. Our methods here will be similar.

Theorem 20. *Let $p = n^{-1/4}$ and $G \in \mathcal{G}(n,p)$. Then whp*

$$\log^* n - \log^* \log^* n - 1 \leq D(G) \leq D_3(G) \leq \log^* n + O(1).$$

The lower bound is very general. We use only the simple fact that any particular unlabeled graph is the value of $\mathcal{G}(n, n^{-1/4})$ with probability at most $\exp\left(-(1/2 - o(1))n^{7/4}\right)$. Let $F(k)$ be the number of pairwise inequivalent sentences about graphs of depth at most k. Then $\Pr[D(G) \leq k] \leq F(k) \exp\left(-(1/2 - o(1))n^{7/4}\right)$ as there are at most $F(k)$ such graphs. From general principles [21, Theorem 2.2.2] we know that $F(k) \leq \mathrm{TOWER}(k + 2 + \log^* k)$. If $k = \log^* n - \log^* \log^* n - 2$, we have $F(k) \leq 2^n$ and hence $\Pr[D(G) \leq k] = o(1)$.

Now we turn to the main part, bounding $D(G)$ from above. For any set W of vertices let $N(W)$ denote the set of common neighbors of W. When $|W| = 4$, $\Pr[N(W) = \emptyset] = (1 - p^4)^{n-4} \approx \mathrm{e}^{-1}$. We are guided by the idea that $N(W) = \emptyset$ is like a random symmetric 4-ary predicate with probability e^{-1}, which is bounded away from both zero and one.

Let $W = \{v_1, v_2, v_3, v_4\}$ be a set of four vertices. Dependent only on W we define:

- $A = N(W)$, the common neighbors of W;
- B, consisting of those $z \notin W \cup A$ such that z is adjacent to precisely four vertices of A and no other $z' \notin W \cup A$ has exactly the same adjacencies to A.

For $w \notin W \cup A$ let $H_w(A)$ denote the 3-uniform hypergraph on A consisting of those triples T so that there is no $z \notin W \cup A$ adjacent to $T \cup \{w\}$. The condition that $z \notin W \cup A$ is a technical convenience that does not asymptotically affect the $H_w(A)$. If, further, $a \in A$, we let $H_{w,a}(A)$ denote the 2-uniform hypergraph (i.e., graph) of pairs T with $T \cup \{a\} \in H_w(A)$. Further, for distinct $a, b \in A$, we let $H_{w,a,b}(A)$ denote the 1-uniform hypergraph (i.e., set) of elements y with $\{a, b, y\} \in H_w(A)$. For $w \notin W \cup A \cup B$ let $H_w(B)$ denote the 3-uniform hypergraph on B consisting of those triples T so that there is no $z \notin W \cup A \cup B$ adjacent to $T \cup \{w\}$. (Again, the condition $z \notin W \cup A \cup B$ is a technical convenience.) Informally, the idea is that the $H_w(A), H_w(B)$ act like random objects with probability e^{-1}.

Call A *universal* if the $H_v(A)$, $v \notin W \cup A$, range over all 3-uniform hypergraphs on A. Call B *splitting* if the $H_v(B)$, $v \notin W \cup A \cup B$, are all different. As there are $2^{\Theta(m^3)}$ 3-uniform hypergraphs over an m-set a simple counting argument gives that if A is universal we must have $|A| = O(\ln^{1/3} n)$ while if B is splitting we must have $|B| = \Omega(\ln^{1/3} n)$.

Our argument splits into two lemmas.

Lemma 21. *Whp there exists a* 4*-set W such that, with A, B as defined above,*

1. *A is universal;*
2. *B is splitting.*

Lemma 22. *Any graph G on n vertices with the property of Lemma 21 has $D_3(G) \leq \log^* n + O(1)$.*

Note that the proof of Lemma 21 is a random graph argument while the proof of Lemma 22 is a logic argument involving no probability.

Proof of Lemma 21. Set $u = \lfloor \ln^{0.3} n \rfloor$. For any set W of four vertices

$$\Pr[|N(W)| = u] = \Pr[\mathrm{Bin}(n-4, p^4) = u] \approx \mathrm{e}^{-1}/u!.$$

Thus the expected number μ of such W has $\mu \approx \binom{n}{4}\mathrm{e}^{-1}/u!$ which approaches infinity. An elementary second moment calculation gives that the number of such W is $(1 + o(1))\,\mu$ whp. Hence it suffices to show that the expected number of W with A having size u but A, B failing the conditions of Lemma 21 is $o(\mu)$. Fix W and A of size u. It suffices to show that A, B satisfy Lemma 21 whp. The conditioning is only on the adjacencies involving a vertex of W, all other adjacencies remain random.

First we show that A is universal. Let Z be those vertices adjacent to four or more vertices of A. Let Z' consist of vertices with at least one neighbor in Z. Whp every four vertices in the graph have $O(\ln n)$ common neighbors so $|Z|$ is polylog while $|Z'| = n^{3/4+o(1)}$.

For each 3-set $Y \subset A$ let $N^-(Y)$ denote those $v \notin W \cup A \cup Z$ which are adjacent to all vertices of Y. The random variables $|N^-(Y)|$, $Y \in \binom{A}{3}$, are independent, each with

Binomial distribution $\mathrm{Bin}(n-o(n),(1+o(1))\,n^{-3/4})$. The probability that $|N^{-}(Y)| > 2n^{1/4}$ or $|N^{-}(Y)| < \frac{1}{2}n^{1/4}$ is then less than $\exp(-cn^{1/4})$ for a constant c by Chernoff bounds. We only need that this probability is $o(n^{-3})$. Thus with high probability for every 3-set $Y \subset A$ we have $|N^{-}(Y)| \in \left[\frac{1}{2}n^{1/4}, 2n^{1/4}\right]$. Condition on these $N^{-}(Y)$ satisfying these conditions. Let R be the (remaining) vertices, not in W, A, Z, Z' nor any of the $N^{-}(Y)$. For $z \in R$ the adjacencies to the $N^{-}(Y)$ are still random. For such z we have $Y \in H_z(A)$ if and only if z is adjacent to no vertex in $N^{-}(Y)$. (Note that z does not send any edges to Z.) As $|N^{-}(Y)| \le 2n^{1/4}$ the probability that z is adjacent to no vertex of $N^{-}(Y)$ is at least e^{-2}. As $|N^{-}(Y)| \ge \frac{1}{2}n^{1/4}$ the probability that z is adjacent to some vertex of $N^{-}(Y)$ is at least $1-\mathrm{e}^{-1/2}$. Set $\gamma = \min(\mathrm{e}^{-2}, 1-\mathrm{e}^{-1/2})$. Then for any hypergraph H on A we have $\Pr[H_z(A) = H] \ge \gamma^{\binom{u}{3}}$. But these events are now independent over the $(1-o(1))\,n$ values $z \in R$ so that the probability that no $H_z(A) = H$ is less than $(1-\gamma^{\binom{u}{3}})^n$. Here because $u^3 = o(\ln n)$ this quantity is less than, say, $\exp(-n^{0.99})$. There are fewer than 2^{u^3} hypergraphs H on A. Hence the probability that any such H is not one of the $H_z(A)$ is less than $2^{u^3}\exp(-n^{0.99})$. The 2^{u^3} term is basically negligible, and the probability that A is not universal is less than $\exp(-n^{0.98})$ and certainly $o(1)$. We note that A being universal will not be fully needed in Lemma 22, we shall need only seven particular values of $H_z(A)$.

Now we look at the size of B. For each $z \notin A \cup W$ the probability that z is adjacent to precisely four elements of A is $\binom{u}{4}p^4(1-p)^{u-4} \approx u^4n^{-1}/24$ and given this the probability that no other z' has the same adjacencies is approximately $(1-p^4)^n \approx \mathrm{e}^{-1}$ so B has expected size $\mu \approx u^4/24\mathrm{e}$. A second moment calculation gives that whp $|B| \approx \mu = \Theta(\ln^{1.2} n)$.

Finally we show that B is splitting. At this stage W, A, B are fixed and all of the adjacencies that do not have at least one vertex from $W \cup A$ are random. Whp no $z \notin W \cup A \cup B$ is adjacent to five (or more) vertices of B. Let Z be those $z \notin W \cup A \cup B$ adjacent to four vertices of B. Whp $|Z|$ is polylog.

For each 3-set $Y \subset B$ we let $N^{*}(Y)$ denote those $v \notin W \cup A \cup B$ which are adjacent to all vertices of Y and $N^{-}(Y) = N^{*}(Y) - Z$. As before, whp all $|N^{*}(Y)|$ have size between $\frac{1}{2}n^{1/4}$ and $2n^{1/4}$ and so the same, asymptotically, holds for the $|N^{-}(Y)|$. As before, we fix the $N^{*}(Y)$ and their adjacencies to Y. Consider distinct $u, u' \notin W \cup A \cup B$. The probability that either u or u' is adjacent to nine (or more) vertices of Z is $o(n^{-2})$. Call a 3-set $Y \subset B$ *exceptional* if u or u' is adjacent to some $z \in Z$ which is adjacent to all of Y. With probability $1-o(n^{-2})$ there are at most $2\cdot 8\cdot 4 = 64$ exceptional Y. Hence the number of non-exceptional Y is $\approx \binom{|B|}{3}$. For the non-exceptional Y we have $Y \in H_u(B)$ if and only if u is adjacent to no vertex in $N^{-}(Y)$ and similarly for $Y \in H_{u'}(B)$. Thus $\Pr[Y \in H_u(B)] \in [\gamma, 1-\gamma]$ with γ as previously defined. Further these events are independent over different non-exceptional Y. Set $\gamma^{*} = \gamma^2 + (1-\gamma)^2$. Then $H_u(B), H_{u'}(B)$ agree on a non-exceptional Y with probability at most γ^{*}. Independence gives that they agree on all non-exceptional Y with probability at most γ^{*} to the $\approx \binom{|B|}{3}$ power. As this power is $\gg \ln n$, the probability is certainly $o(n^{-2})$. There are $O(n^2)$ choices of u, u' so whp no $H_u(B) = H_{u'}(B)$. ∎

Proof of Lemma 22. The main portion of the argument consists of placing an arithmetic structure on A in such a way that any vertex in A can be described with quantifier depth $\log^{*}|A| + O(1) \le \log^{*} n + O(1)$. For convenience we assume $|A| = 3s+2$. (Otherwise we would specify one or two extra elements of A by reserving for them one or two special variables and thereby increasing the depth by at most two. Note that the set A with these elements removed stays universal.) Label the elements of A by $a, b, x_1, \dots, x_s, y_1, \dots, y_s, z_1, \dots, z_s$ in an arbitrary way. Now we, effectively, model arithmetic on A. From the universality there exist $w_1, \dots, w_7$ (witnesses) such that:

1. H_{w_1} consists of the triples $\{x_i, y_i, z_i\}$.
2. $H_{w_2,a,b}$ consists of the elements x_i.
3. $H_{w_3,a,b}$ consists of the elements y_i.
4. $H_{w_4,a}$ consists of all pairs $\{x_i, y_j\}$ with $i \leq j$.
5. H_{w_5} consists of all triples $\{x_i, y_j, z_{i+j}\}$ with $1 \leq i,j, i+j \leq s$.
6. H_{w_6} consists of all triples $\{x_i, y_j, z_{i \cdot j}\}$ with $1 \leq i,j, i \cdot j \leq s$.
7. $H_{w_7,a}$ consists of all pairs $\{x_i, y_{2^i}\}$ with $1 \leq i \leq s$ and $2^i \leq s$.

Using 13 first-order variables to denote the v_1, v_2, v_3, v_4 which define A, the special elements $a, b \in A$, and the witnesses $w_1, \ldots, w_7$, we give a first order expression which forces A to have the above form. Note that membership in A is given by a first order statement and membership in an H_w or $H_{w,a}$ or $H_{w,a,b}$ is given by a first order statement in terms of the variables. Let $A^- = A - \{a, b\}$ for convenience. Now we express the following six properties by first-order formulas.

P_1 (1-Factor) H_{w_1} consists of vertex disjoint triples and every element of A^-, and only those elements, are in such a triple.

P_2 (Splitting the 1-Factor) For each triple in H_{w_1} exactly one of the elements is in $H_{w_2,a,b}$ and exactly one (a different one) is in $H_{w_3,a,b}$. Now let (for convenience) X denote those $x \in A^-$ with $x \in H_{w_2,a,b}$, and let Y denote those $y \in A^-$ with $y \in H_{w_3,a,b}$ and Z the other elements of A^-. Henceforth the use of the letter x, y, z shall tacitly assume that the element is in the respective set X, Y, Z. We write $x \leftrightarrow y$ or $x \leftrightarrow z$ or $y \leftrightarrow z$ if the two elements are in a common triple in H_{w_1}.

P_3 (Creating $\leq$) Here adjacency is in $H_{w_4,a}$. We require that all adjacencies be between an x and a y. Let $N(x)$ denote the y adjacent to x. We require that for every x, x' either $N(x) \subset N(x')$ or $N(x') \subset N(x)$ with equality only when $x = x'$. We require that when $y \leftrightarrow x$, then $y \in N(x)$. This forces the $N(x)$ to form a chain and so the x and y can be renumbered to fit the condition. We now define $x \leq x'$ by $N(x') \subseteq N(x)$. The relations $\geq, >, <$ have their natural Boolean meaning in terms of $\leq$. We define $y \leq y'$ and $z \leq z'$ by $x \leq x'$, where $x \leftrightarrow y \leftrightarrow z$ and $x' \leftrightarrow y' \leftrightarrow z'$. We let x_1, y_1, z_1 denote the first elements under $\leq$ and x_s, y_s, z_s the last elements. The notions of successor x^+ and predecessor x^- are naturally defined (when they exist) in terms of $\leq$. We let y_2 and z_2 denote the successors of y_1 and z_1 respectively.

P_4 (Creating addition) Addition is generated from the formulas $\alpha + 1 = \alpha^+$ and $\alpha + \beta^+ = (\alpha + \beta)^+$, though we need some care as addition in this model is not always defined. For every $x \in X, y \in Y$ there is at most one $z \in Z$ with $\{x, y, z\} \in H_{w_5}$. $\{x, y_1, z\} \in H_{w_5}$ if and only if $x \neq x_s$ and $z \leftrightarrow x^+$. If $y \neq y_s$, then $\{x, y^+, z\} \in H_{w_5}$ if and only if $z \neq z_1$ and $\{x, y, z^-\} \in H_{w_5}$. We let $x + x' = x^*$ denote that $\{x, y', z^*\} \in H_{w_5}$, where $y' \leftrightarrow x'$ and $y^* \leftrightarrow x^*$. Let $x + z = z'$ mean that when z, z' are replaced by their $\leftrightarrow$ elements in x that then we have the equality, and similarly for other forms like $y + y' = z$.

P_5 (Creating multiplication) Multiplication is generated from the formulas $\alpha \cdot 1 = \alpha$ and $\alpha \cdot \beta^+ = (\alpha \cdot \beta) + \beta$, though we need some care as addition in this model is not always defined. For every $x \in X, y \in Y$ there is at most one $z \in Z$ with $\{x, y, z\} \in H_{w_6}$. $\{x, y_1, z\} \in H_{w_6}$ precisely when $x \leftrightarrow z$. If $y \neq y_s$, then $\{x, y^+, z\} \in H_{w_6}$ if and only if $\{x, y, z'\} \in H_{w_6}$ and $x + z' = z$ for some z'.

P_6 (Creating exponentiation) Base two exponentiation is defined by $2^1 = 2$ and $2^{\alpha^+} = 2^\alpha + 2^\alpha$, though we need some care as addition in this model is not always defined. For every $x \in X$ there is at most one $y \in Y$ with $\{x, y\} \in H_{w_7,a}$. $\{x_1, y_2\} \in H_{w_7,a}$. If $x \neq x_s$, then $\{x^+, y\} \in H_{w_7,a}$ if and only if $\{x, y'\} \in H_{w_7,a}$ and $y' + y' = y$ for some y'. We write $x' = 2^x$ if $\{x, y'\} \in H_{w_7,a}$ and $x' \leftrightarrow y'$.

In outline, the defining formula is constructed in the following way. An integer x can be described with quantifier depth $O(\log^* n)$ by listing its binary digits, where in order to specify the position d of each digit we apply recursion on $d \leq \log_2 x$. Now the elements v of B are identified by describing the four vertices of A that v is adjacent to. Any $v \notin W \cup A \cup B$ is described by listing the edges of $H_v(B)$. The assumption that B is splitting means that we have identified all the vertices; now we can just list all the graph adjacencies. Also, we have to be careful so that the constructed defining sentence has the alteration rank at most 3, as claimed.

We will say that a sequence of symbols $\forall$ and $\exists$ has form $\forall^*\exists^*\forall^*$ if it consists of a block of all $\forall$ followed by a block of all $\exists$ and a concluding block of all $\forall$ (some blocks may be empty). Any other regular expression over alphabet $\{\forall, \exists\}$, as $\exists^*\forall^*$ or $\exists^*\forall^*\exists^*$, will be understood similarly. Furthermore, we will say that a first order formula P is a $\forall^*\exists^*\forall^*$ formula if any sequence of nested quantifiers in P has form $\forall^*\exists^*\forall^*$.

Each formula P_i, $i \leq 6$, has free variables in the set $\{v_1, \ldots, v_4, w_1, \ldots, w_7, a, b\}$ (for the simplicity of notation we use the same letters for variables and vertices corresponding to each other). A straightforward inspection shows that every P_i is a $\forall^*\exists^*\forall^*$ formula. As all P_i's are fixed formulas, they have constant quantifier depth.

For each x_i in X, $1 \leq i \leq s$, we now construct a formula $A_{x_i}(x)$ that has one free variable x and describes x_i; i.e., $A_{x_i}(x)$ will be true for $x = x_i$ and false for any other vertex of G. Using the arithmetics introduced by $P_1, \ldots, P_6$, we identify a vertex x_i with the integer i. We will need a formula $D(d, x)$ whose truth value coincides with the dth binary digit of x, if d and x range in X and $d \leq \lfloor \log_2 x \rfloor$ (we count the first digit as the first on the right, zero if and only if x is even). The 1st digit of x is zero if and only if $x = x' + x'$ for some x'. Otherwise the dth digit is zero if and only if there exist q, r $x = q \cdot 2^d + r$ with $r < 2^{d-1}$ or if $x = q \cdot 2^d$ or if $x < 2^{d-1}$. (This technical complication is caused by leaving zero out of the model.) This is first order as we already have multiplication, exponentiation, less than, and addition. $D(d, x)$ can be written so that any sequence of nested quantifiers in this formula has form $\exists^*\forall^*$ or $\forall^*\exists^*$.

The formula $A_{x_1}(x)$ just says that $x \in X$ and $x \leq x'$ for all x' in X. For $i > 1$, the construction of $A_{x_i}(x)$ is recursive. First of all, we start $A_{x_i}(x)$ with the assertion that $x \in X$. Let x_i have m digits. If $2^m \leq s$, we put in $A_{x_i}(x)$ the assertion $\exists_{x'}(A_m(x') \wedge x \leq 2^{x'})$, where $A_m(x)$ is constructed recursively. The same can as well be expressed as $\forall_{x'}(\neg A_m(x') \vee x \leq 2^{x'})$. Note that $x \leq 2^{x'}$ is an $\exists^*\forall^*$ formula. If $2^m > s$, we reset $m = \lfloor \log s \rfloor$ and put no assertion in $A_{x_i}(x)$ at this stage. Furthermore, for every $j \leq m$, we put in $A_{x_i}(x)$ the assertion $\exists_{x'}(A_{x_j}(x') \wedge D(x', x))$ or its $\forall$-version, if the jth digit of x_i is 1. If the jth digit of x_i is 0, we put the assertion $\exists_{x'}(A_{x_j}(x') \wedge \neg D(x', x))$ or its $\forall$-version. Here $A_{x_j}(x)$ is constructed recursively. Therewith $A_{x_i}(x)$ is completely specified, with the freedom to choose front quantifiers. Observe that $A_{x_i}(x)$ has subformulas $A_m(x)$ and $A_{x_j}(x)$ only with $x_j \leq m \leq \lfloor \log x_i \rfloor + 1$. It easily follows that every A_{x_i} has quantifier depth at most $\log^* s + O(1)$.

Choosing appropriately $\exists$- or $\forall$-variants for ingredients of A_{x_i} and all subsequent $A_m(x)$ and $A_{x_j}(x)$, we can write A_{x_i} in either of two forms, $\exists^*\forall^*\exists^*$ and $\forall^*\exists^*\forall^*$.

An element y_i of Y is described by formula $A_{y_i}(y)$ which has two variants, $y \in Y \wedge \exists_x(A_{x_i}(x) \wedge x \leftrightarrow y)$ or $y \in Y \wedge \forall_x(\neg A_{x_i}(x) \vee x \leftrightarrow y)$. Elements of Z are described similarly. Note that $x \leftrightarrow y$ is an $\exists^*\forall^*$ formula.

The distinguished vertices are described trivially. For example, $A_a(x)$ is just $x = a$. Thus, all elements in $A \cup W \cup \{w_1, \ldots, w_7\}$ have first order descriptions.

Note that the statement $x \in B$ is expressible by a $\forall^*\exists^*$ formula. A vertex x in B with neighbors a_1, a_2, a_3, a_4 in A is described by a formula $A_v(x)$ that is $x \in B \wedge \bigwedge_{i=1}^4 \exists_y(A_{a_i}(y) \wedge x \sim y)$ or the $\forall$-variant thereof. (Recall that $\sim$ stands for the graph adjacency relation.)

For any remaining vertex v, its description $A_v(x)$ consists of the assertion $x \notin W \cup A \cup B \cup \{w_1, \ldots, w_7\}$ and the conjunction of $\exists_{y_1,y_2,y_3}(A_{b_1}(y_1) \wedge A_{b_2}(y_2) \wedge A_{b_3}(y_3) \wedge \{y_1, y_2, y_3\} \in H_x(B))$ over all triples $\{b_1, b_2, b_3\} \in H_x(B)$ and $\exists_{y_1,y_2,y_3}(A_{b_1}(y_1) \wedge A_{b_2}(y_2) \wedge A_{b_3}(y_3) \wedge \{y_1, y_2, y_3\} \notin H_x(B))$ over all triples $\{b_1, b_2, b_3\} \notin H_x(B)$, $b_1, b_2, b_3 \in B$. Again note that the conjunctive members have alternative $\forall$-variants. The assumption that B is splitting means that the constructed $A_v(x)$ indeed describes the v.

Summarizing, we conclude that every vertex v of G is described by a first-order formula $A_v(x)$ that has quantifier depth $\log^* n + O(1)$ and can be written in either of two forms, $\exists^*\forall^*\exists^*$ and $\forall^*\exists^*\forall^*$.

We now give a first-order sentence defining G. It is

$$\exists_{v_1,\ldots,v_4,w_1,\ldots,w_7,a,b}\left(\bigwedge_{i=1}^{6} P_i \wedge Q \wedge V \wedge E\right), \tag{29}$$

where $P_1, \ldots, P_6$ are specified above and Q, V, E are as follows. The formula Q says that all $v_1, \ldots, v_4, w_1, \ldots, w_7, a, b$ are pairwise distinct, no w_i is in $W \cup A$, and a, b are in A. The formula V, specifying the vertex set, is this:

$$\bigwedge_{v \in V(G)} \left(\exists_x A_v(x) \wedge \forall_{x_1,x_2}(A_v(x_1) \wedge A_v(x_2) \to x_1 = x_2)\right) \wedge \forall_x \bigvee_{v \in V(G)} A_v(x).$$

Note that V determines, in particular, the number s (and the size of A). Finally, the formula E, specifying all (non)adjacencies between the vertices, is as follows:

$$\forall_{x_1,x_2}\left(\bigwedge_{\{u,v\} \in E(G)} (A_u(x_1) \wedge A_v(x_2) \to x_1 \sim x_2) \wedge \bigwedge_{\{u,v\} \notin E(G)} (A_u(x_1) \wedge A_v(x_2) \to \neg(x_1 \sim x_2))\right).$$

It is easy to see that the defining formula (29) has quantifier depth $\log^* n + O(1)$. By appropriate choice of $\exists^*\forall^*\exists^*$- or $\forall^*\exists^*\forall^*$-form for each occurrence of A_v, one can ensure that (29) is an $\exists^*\forall^*\exists^*\forall^*$ formula and hence has alternation number 3. ∎

7. CONCLUDING REMARKS

Our Theorem 2 has a strong link with the zero-one law which was discovered independently by Glebskii et al. [11] and Fagin [9] and says that $G \in \mathcal{G}(n, \frac{1}{2})$ satisfies any fixed first order sentence with probability approaching either 0 or 1. Given $\epsilon \in (0, 1)$, define $T_\epsilon(n)$ to be the maximum k such that, if $n_1, n_2 \geq n$, then $D(G, H) > k$ with probability at least $1 - \epsilon$ for independent $G \in \mathcal{G}(n_1, p)$ and $H \in \mathcal{G}(n_2, p)$. The Bridge Theorem [21, Theorem 2.5.1] says that, in a rather general setting, a zero-one law is obeyed iff, for each ϵ, $T_\epsilon(n)$ tends to the infinity as n increases. Spencer and St. John [22] call $T_\epsilon(n)$ the *tenacity function* and suggest it as a quantitative measure for observation of a zero-one law. While in [22] the tenacity function is studied for words, here we are able to find its asymptotics in the case of graphs. Using Theorem 2 and noting that the lower bound based on the k-extension property goes through for $D(G, H)$ with both G and H random, we obtain $T_\epsilon(n) = \log_{1/p} n + O(\ln \ln n)$, irrespective of the constant ϵ.

Another interesting first-order parameter of a graph G is $I(G)$, the smallest depth of a sentence distinguishing G from any non-isomorphic graph of the *same* order as G. Of course, $I(G) \leq D(G)$, so all upper bounds we have proved apply to $I(G)$ as well. All our lower bounds also apply to $I(G)$ with the exceptions of Theorem 19. Its $I(G)$-analog would say that, for $G \in \mathcal{G}(n, \frac{c}{n})$ with $c < c_0$, where $c_0 = 1.034\ldots$, we have whp

$$I(G) = (1 + o(1))t_2(G) = (c\mathrm{e}^{-2c}/2 + o(1))n, \tag{30}$$

where $t_2(G)$ denotes the number of isolated edges. The reason is that if $G \ncong G'$ but $v(G) = v(G')$, then the multiplicities of at least two non-isomorphic components must differ while the two most frequent components in G are isolated vertices and edges. (And the order of the giant component catches up with $t_2(G)$ at $p \approx \frac{c_0}{n}$.) The Reader should not have any problem in filling up the missing details.

We make the following general conjecture.

Conjecture 23. *Let $\varepsilon > 0$ be fixed and $n^{-1+\varepsilon} \leq p \leq \frac{1}{2}$. Then whp $D(G) = O(\ln n)$.*

One can also ask about $D_{\#}$, the analog of $D(G)$ when we add counting to first order logic. Here, the situation is strikingly different. A result of Babai and Kučera [2] (combined with Immerman and Lander's [12] logical characterization of the vertex refinement step in [2]) implies that whp $G \in \mathcal{G}(n, \frac{1}{2})$ can be defined by a first-order sentence with counting of quantifier depth at most 4.

REFERENCES

[1] N. Alon and J. Spencer, The probabilistic method, 2nd edition, Wiley Interscience, New York, 2000.

[2] L. Babai and L. Kučera, Canonical labeling of graphs in linear average time, Proc 20th IEEE Symp Foundations of Computer Science, 1979, pp. 39–46.

[3] A. D. Barbour, Poisson convergence and random graphs, Math Proc Cambridge Philos Soc 92 (1982), 349–359.

[4] T. Bohman, A. Frieze, T. Łuczak, O. Pikhurko, C. Smyth, J. H. Spencer, and O. Verbitsky, First order definability of trees and sparse random structures, in preparation, 2004.

[5] B. Bollobás, Random graphs, 2nd edition, Cambridge University Press, Cambridge, 2001.

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

[7] K. J. Compton, A logical approach to asymptotic combinatorics. I. First order properties, Adv Math 65 (1987), 65–96.

[8] A. Ehrenfeucht, An application of games to the completeness problem for formalized theories, Fund Math 49 (1961), 129–41.

[9] R. Fagin, Probabilities on finite models, J Symbolic Logic 41 (1976), 50–58.

[10] R. Fraïssé, Sur quelques classifications des systèmes de relations, Publ Sci Univ Alger 1 (1954), 35–182.

[11] Y. V. Glebskii, D. I. Kogan, M. I. Liogonkii, and V. A. Talanov, Volume and fraction of satisfiability of formulas of the lower predicate calculus, Kibernetyka (Kyiv) (2) (1969), 17–27.

[12] N. Immerman and E. Lander, "Describing graphs: A first-order approach to graph canonization," Complexity theory retrospective, Ed. A. Selman, pp. 59–81. Springer, New York, 1990.

[13] Ph. G. Kolaitis, H. J. Prömel, and B. L. Rothschild, Asymptotic enumeration and a 0-1 law for *m*-clique free graphs, Bull Amer Math Soc 13 (1985), 160–162.

[14] Ph. G. Kolaitis, H. J. Prömel, and B. L. Rothschild, K_{l+1}-free graphs: Asymptotic structure and a 0-1 law, Trans Amer Math Soc 303 (1987), 637–671.

[15] T. Łuczak and J. Spencer, When does the zero-one law hold?, J Amer Math Soc 4 (1991), 451–468.

[16] O. Pikhurko, J. Spencer, and O. Verbitsky, Succinct definitions in the first order graph theory, submitted, E-print `arXiv:math.LO/0405326`, 2004.

[17] O. Pikhurko, H. Veith, and O. Verbitsky, The first order definability of graphs: Upper bounds for quantifier ranks, submitted, E-print `arXiv:math.CO/0311041`, 2003.

[18] O. Pikhurko and O. Verbitsky, Descriptive complexity of finite structures: Saving the quantifier rank, To appear in J Symbolic Logic, E-print `arXiv:math.LO/0305244`, 2003.

[19] S. Shelah, Very weak zero one law for random graphs with order and random binary functions, Random Structures Algorithms 9 (1996), 351–358.

[20] S. Shelah and J. Spencer, Zero-one laws for sparse random graphs, J Amer Math Soc 1 (1988), 97–115.

[21] J. Spencer, The strange logic of random graphs, Springer, New York, 2001.

[22] J. Spencer and K. St. John, The tenacity of zero-one laws, Electron J Combin 8 (2001), 14 pp.