986 revision:2015-10-21_2 modified:2015-10-21

Random graphs and Lindström quantifiers for natural graph properties

Simi Haber Saharon Shelah

Abstract

We study zero-one laws for random graphs. We focus on the following question that was asked by many: Given a graph property P, is there a language of graphs able to express P while obeying the zero-one law? Our results show that on the one hand there is a (regular) language able to express connectivity and k-colorability for any constant k and still obey the zero-one law. On the other hand we show that in any (semiregular) language strong enough to express Hamiltonicity one can interpret arithmetic and thus the zero-one law fails miserably. This answers a question of Blass and Harary.

1 Introduction

1.1 Motivation and definitions

In this paper we study questions related to the following observation: Natural graph properties are either held by most graphs or not held by most graphs.

The *binomial random graph* $G(n,p)$ is a probability distribution over all labeled graphs of order n. For convenience we assume that the vertex set is simply the set $\{1,2,\dots,n\}$. Sampling from $G(n,p)$ is done by including every possible edge in the graph at random with probability p independently of all other edges. Random graphs were extensively studied from the sixties, starting with the seminal work of Erdős and Rényi [13, 14]. From the beginning of the study of random graphs a phenomenon was spotted. Consider the simplest model of random graph, $G(n,1/2)$, in which we pick a graph uniformly at random from all the labeled graphs of n vertices. For many natural graphs properties — properties like connectivity, containing a Hamilton path or cycle, not being colorable in some fixed number of colors, not being planar and containing a small fixed graph as a subgraph — $G(n, 1/2)$ *asymptotically almost surely* (abbreviated a.a.s.) has these properties. This means that the probability for which a random graph has any one of these properties tends to one as the size of the graph tends to infinity. We may remove the negation before colorability and planarity and say that for any of these properties either a.a.s. $G(n, 1/2)$ has these property or a.a.s. $G(n, 1/2)$ does not have it.

The phenomenon remains valid if we replace $1/2$ by p for any constant $0 < p < 1$, so from now on p will be some fixed constant real number (strictly between zero and one) representing the edge probability of the random graph.

This phenomenon suggests the following definition:

Definition 1. Let $\mathcal{A}$ be a set of graph properties. We say that $\mathcal{A}$ *obeys the zero-one law* if for every property $\varphi \in \mathcal{A}$ one has

$$\lim_{n\to\infty} \Pr[G(n,p) \text{ has } \varphi] \in \{0,1\}.$$

Having this definition we may rephrase the first sentence of this introduction by saying that if $\mathcal{A}$ is a set of natural graph properties then $\mathcal{A}$ obeys the zero-one law. But what are "natural graph properties"? A natural interpretation is: all properties defined by a sentence in a language $\mathcal{L}$. Indeed, the first zero-one law is of this sort[1] where the language was $\mathcal{FO}$, the first order language (Glebskiĭ et al. in [16] and Fagin in [15], see below). Unfortunately, most classical graph theoretic properties are not expressible in $\mathcal{FO}$. A reasonable list of the most extensively studied graph theoretic properties should include connectivity, k-colorability, Hamiltonicity and also planarity. Thus we suggest to consider the language for graphs obtainable from first order logic strengthen by a property from this list. This is done by adding a quantifier for the property, as detailed below.

Considering graphs, the most basic formal language is the first order language of graphs denoted here by $\mathcal{FO}$. It consists of the following symbols:

1. Variables, denoted along this paper by lower case Latin letters x, y, z. In $\mathcal{FO}$, variables stand for vertices solely.

[1] In fact, to the best of our knowledge, in *all* of the zero-one laws for graphs the set of properties is the set of sentences in some formal language.

986 revision:2015-10-21_2 modified:2015-10-21

2. Relations. There are exactly two of these: adjacency, denoted here by $\sim$, and equality, denoted as usual by $=$. Thus it is possible to write $x = y$ or $y \sim z$. The relations is what makes this language a language of graphs.

3. Quantifiers. Again, there are two, the existential $\exists$ and the universal $\forall$. These can be applied only on variables which means that quantification in $\mathcal{FO}$ is only possible over vertices.

4. Boolean connectives, like $\wedge, \vee, \neg$ and $\rightarrow$.

Notice that there are no constants and no functions in this language. As usual, we shall also use parentheses and punctuation marks for the benefit of readability.

For example, in $\mathcal{FO}$ we may write

$$\exists x \exists y \exists z.\, [\neg(x = y) \wedge \neg(x = z) \wedge \neg(y = z) \wedge x \sim y \wedge x \sim z \wedge y \sim z],$$

which means "there exists a triangle in the graph". Another example might be "there are no isolated vertices":

$$\forall x \exists y.\, [\neg(x = y) \wedge x \sim y].$$

There are few limitations on the language. Every formula must be of finite length, and again, variables stand only for vertices so in particular we may only quantify over vertices. This is a crucial difference between first order and higher order logics.

In the *second order* language of graphs, denoted here by $\mathcal{SO}$, we have *relational variables* (also called second order variables). Adhering customary notation, we denote relational variables by capital Latin letters. We write the arity of a second order variable in superscript near the binding operator. If A is an unary relation, we use $x \in A$ instead of $A(x)$, and if B is a binary relation we may write xBy instead of $B(x, y)$. Here are a few examples of sentences in $\mathcal{SO}$ representing graph properties:

Connectivity A graph is connected if there is a walk or a path[2] between any two vertices. Equivalently, there is an edge between the parts of

[2] A *walk* in a graph $G = (V, E)$ is a sequence of vertices $v_0, v_1, \ldots, v_l$ such that for any $1 \leq i \leq l$ one has $v_{i-1} \sim v_i$. If additionally the vertices are all distinct, it is called a *path*. A *cycle* is a walk in which the first and last vertices are identical, and all others are distinct.

modified:2015-10-21
986 revision:2015-10-21_2

any nontrivial partition:

$$\forall A^1. [((\exists x. x \in A) \land (\exists x. x \notin A)) \to (\exists x \in A, y \notin A. x \sim y)].$$

3-colorability A graph is k-colorable if there is a partition of the vertex set into k parts (colors), such that there is no edge between ny vertices in the same part. This may be written as:

$$\begin{aligned}\exists R^1, G^1, B^1. [&\mathrm{Partition}(R, G, B) \land \\ &\land (\forall x, y. (x, y \in R) \to x \not\sim y) \land \\ &\land (\forall x, y. (x, y \in G) \to x \not\sim y) \land \\ &\land (\forall x, y. (x, y \in B) \to x \not\sim y)],\end{aligned}$$

where Partition(R, G, B) is a first order formula saying that the sets R, G and B are mutually disjoint and their union equals the set of vertices.

Hamiltonicity A Hamilton path in a graph is a path that visits every vertex exactly once. A Hamilton cycle is a Hamilton path where the first and last vertices are adjacent. We say that a graph is Hamiltonian if it contains a Hamilton cycle.

In the following sentence Min$(<, x)$ is a first order formula expressing the fact that x is a minimal element with respect to an order $<$ (and similarly for Max$(<, y)$). The first three lines say that $<$ is a linear order, the fourth line says that each vertex is adjacent to its successor in that order and the last line adds that the first vertex is adjacent to the last one.

$$\begin{aligned}\exists <^2. [\, &(\forall x. \neg(x < x)) \land \\ &(\forall x, y. ((x < y) \lor (y < x)) \land \neg((x < y) \land (y < x))) \land \\ &(\forall x, y, z. ((x < y) \land (y < z)) \to (x < z)) \land \\ &(\forall x, y. (\neg\exists z.(x < z) \land (z < y)) \to x \sim y) \land \\ &\exists x, y. \mathrm{Min}(<, x) \land \mathrm{Max}(<, y) \land y \sim x].\end{aligned}$$

1.2 Previous results

The first zero-one law was proven by Glebskiĭ et al. in [16] and independently by Fagin [15]. They showed that the set of properties describable in $\mathcal{FO}$ obeys

modified:2015-10-21

986 revision:2015-10-21_2

the zero-one law. In his paper Fagin demonstrated that $\mathcal{SO}$ does not obey the zero-law by expressing the sentence "the number of vertices is even":

$$\exists P^2.\, [\forall x.\, \exists! y.\, (x \neq y \wedge xPy \wedge yPx)]\,. \tag{1}$$

The quantifier $\exists!$ stands for there exists exactly one element such that... Notice that parity is a property of a set (of the set of vertices in our case) and thus adjacency does not appear in this sentence. When the zero-one law fails for $\mathcal{A}$, we may argue for less.

Definition 2. We say that $\mathcal{A}$ obeys a *limit law* if for every property $p \in \mathcal{A}$ the limit

$$\lim_{n\to\infty} \Pr[G(n,p) \text{ has } P]$$

exists.

Clearly, the sentence in Equation 1 demonstrates that $\mathcal{SO}$ does not even obey the limit law.

At this point it is worth mentioning that $\mathcal{FO}$ is rather weak. Of the list of "natural" properties above — connectivity, Hamiltonicity, k-colorability (for $k \geq 2$), planarity and having a fixed graph as a subgraph — only the last is a first order property. Thus it is clear that the zero-one law of Fagin and Glebskiĭ et al. does not capture the phenomenon aforementioned. On the other hand, as Fagin showed, $\mathcal{SO}$ is too expressive. There are many papers dealing with the problem of finding a language that is strong enough to express some graph properties that are not first order expressible on the one hand, while still obeying the zero-one law on the other. In the following we will mention some of these results. The following description is far from being comprehensive. For a survey of the results in this field see, e.g., [9].

Looking at examples such as above (in particular the sentences representing connectivity and k-colorability), it seems reasonable to study the expressive power of the *monadic second order* logic. In the monadic second order language of graphs, denoted here by $\mathcal{MSO}$, all second order variables must be of arity one, that is, all second order variables represent sets. It was asked by Blass and Harary in 1979 [5] whether $\mathcal{MSO}$ obeys the zero-one law. In their 1985 paper [22], Kaufmann and Shelah provided a strong negative answer — even a fraction of $\mathcal{MSO}$ is enough to express properties for which the asymptotic probability does not exist. Let $\mathcal{MESO}$ be the set of formulas having the following structure: $\exists \bar{A}.\, \varphi(\bar{A}, \bar{a})$, where $\bar{A}$ is a vector of

986 revision:2015-10-21_2 modified:2015-10-21

unary second order variables, $\bar{x}$ are first order variables and φ is a first order formula. That is, in $\mathcal{MESO}$ we may only existentially quantify over sets and at the beginning of the formula. The 3-colorability sentence in the second example above is a $\mathcal{MESO}$ sentence. Notice that $\mathcal{MESO}$ is not closed under negation. In particular, connectivity is not $\mathcal{MESO}$-expressible, while being disconnected obviously is. Kaufmann and Shelah showed that one can interpret a segment of arithmetic in $\mathcal{MESO}$, and hence express properties like, say, $0 \leq \sqrt{n} \leq 4 \pmod{10}$. Clearly the last property has no asymptotic probability; moreover, the limit superior of its probability sequence is one and the limit inferior is zero.

In order to give the full strength formulation of the result of Kaufmann and Shelah let us define the notion of *arithmetization*. The language of arithmetic is the first order language with universe set $\mathbb{N}$ and vocabulary $\{<, +, \cdot\}$ where $<$ is the natural order of integers and $+, \cdot$ have their usual meaning (see, e.g., [12]).

Definition 3. Let $\mathcal{L}$ be a language for the class of graphs and let $\mathcal{G} = (G_n)$ be a sequence of probability distributions over graphs of order n. We say that *the pair* $(\mathcal{L}, \mathcal{G})$ *can interpret arithmetic for a function* $f\colon \mathbb{N} \to \mathbb{N}$ if for every sentence φ in the language of arithmetic there is a sentence $\psi_\varphi \in \mathcal{L}$ such that

$$\lim_{n\to\infty} \Pr\left[G_n \models \psi_\varphi \iff \mathbb{N}|_{f(n)} \models \varphi\right] = 1.$$

If the pair $(\mathcal{L}, \mathcal{G})$ can interpret arithmetic for some function $f\colon \mathbb{N} \to \mathbb{N}$ we say that *the pair* $(\mathcal{L}, \mathcal{G})$ *has arithmetization.*

When a pair $(\mathcal{L}, \mathcal{G})$ has arithmetization it is, in a sense, the farthest that can be from obeying a zero-one law. For example, in this situation for any recursive real $\alpha \in [0, 1]$ there is a sentence $\varphi_\alpha \in \mathcal{L}$ having asymptotic probability α. The aforementioned result of Kaufmann and Shelah about $\mathcal{MESO}$ is of this sort, demonstrating that $(\mathcal{MESO}, G(n, 1/2))$ has arithmetization. There are situations in which $\mathcal{FO}$ has arithmetization. Let $\alpha \in (0, 1)$ be rational, then the pairs $(\mathcal{FO}, G(n, p = n^\alpha))$ and $(\mathcal{FO}, G_{n,d=n^{1-\alpha}})$ have arithmetization (where $G_{n,d=n^{1-\alpha}}$ is the *random regular graph* with degree d. See [29] and [17] respectively). Also, in [20] it is showed that if $d \geq 2$ is a constant integer and r is a small enough constant then the pair $(\mathcal{FO}, G(n; \mathbb{T}^d, r))$ has arithmetization (where $\mathbb{T}^d$ is the d-dimensional torus and $G(n; \mathbb{T}^d, r)$ is the *random geometric graph* with distance parameter r).

986 revision:2015-10-21_2 modified:2015-10-21

In view of the last result it seems reasonable to look for less expressive fragments of second order logic for which the zero-one law will hold. In their 1987 paper [24] Kolaitis and Vardi proved a zero-one law for the *strict* Σ_1^1 language — the set of sentences of the form $\exists \bar{S}.\, \psi(\bar{S})$ where ψ is from the *Bernays-Schönfinkel* class. That is, the set of all sentences of the form $\exists \bar{S}.\, \exists \bar{x}.\, \forall \bar{y}.\, \varphi(\bar{S}, \bar{x}, \bar{y})$ where $\bar{S}$ is a vector of second order variables, $\bar{x}$ and $\bar{y}$ are vectors of first order variables and $\varphi(\bar{S}, \bar{x}, \bar{y})$ is a quantifier free formula. Notice that 3-colorability is a strict Σ_1^1 property, as well as disconnectivity. On the other hand connectivity is not a strict Σ_1^1 property. A line of research was started by [24], aiming to characterize the Σ_1^1 fragments defined by first-order prefix classes according to adherence to the zero-one law. The classification was completed in [3]. See [25] for a survey.

Another family of languages studied in this context is the family of languages one get from adding a recursive operator to the first order logic. These languages can express connectivity but not k-colorability. There are a few such languages known to obey the zero-one law [31, 4].

The last result we shall mention in this section deals with a different strengthening of the first order logic. In [23] Kolaitis and Kopparty considered the language one gets by augmenting a *parity* quantifier to $\mathcal{FO}$. Their result was a *modular* limit-law:

Theorem 4 ([23])**.** *Let $\mathcal{FO}[\oplus]$ be the regular language obtained by adding parity to the first order language. Then for every property $P \in \mathcal{FO}[\oplus]$ there are two rational numbers α_0, α_1 such that*

$$\lim_{n=2k+i\to\infty} \Pr[G(n,p) \textit{ has } P] = \alpha_i$$

The result generalizes in the natural way to general modulo k operators for any constant k. Theorem 4 means that the fact that a language is able to express parity is actually not that bad. Indeed, parity has no asymptotic probability, but we can think of the graph sequence[3] as two separate sequences, odd and even, and then we do get a limit for the asymptotic probability in each of these sequences. Adding parity clearly lets us say different things for odd and even graphs, Theorem 4 tells us that it does not give more in terms of expressive power.

As mentioned above, there are many other papers along this line. To the best of our knowledge none of these papers presented a formal language

[3]Actually, we have a sequence of probability spaces over graphs.

modified:2015-10-21
986 revision:2015-10-21_2

strong enough to express 3-colorability and connectivity while obeying the zero-one law, The same question for Hamiltonicity was explicitly asked first in [5] and then by many others.

The results in this paper (Theorems 10 and 12) give the following answer to the question above: On the one hand, there is a regular language able to express connectivity and k-colorability for any fixed $k \geq 2$ while obeying the zero-one law. On the other hand, our main result states that any semiregular language able to express Hamiltonicity can express arithmetic as well (and therefore it violates even the modular limit law of Kolaitis and Kopparty). Planarity behaves similarly to connectivity and colorability. A result including planarity will be published elsewhere.

Before stating our theorems we shall present the connected notions of regular languages and Lindström quantifiers.

1.3 Regular languages and Lindström quantifiers

There is a trivial "language" that can express the properties listed above while simultaneously obeying the zero-one law — simply take $\mathcal{FO}$ and add the sentences "G is Hamiltonian", "G is planar" and so on. Clearly this language misses some notion of closure. To avoid such trivialities we need to define what kind of languages are accepted. The definitions in this section follow the ideas of Lindström [26, 27]. Our notation is taken from [11], in which a full treatment of the notions in this section may be found.

A language $\mathcal{L}$ is called *semiregular* if it is closed under "first order operations" and it is also closed under substitution of a formula for a predicate. That is, $\mathcal{L}$ is required to contain the atomic formulas (in our case, formulas of the form $x = y$ and $x \sim y$), to be closed under Boolean connectives (e.g., $\neg$ and $\wedge$) and existential quantification and finally to allow redefinition of the predicates through formulas in $\mathcal{L}$. A language is called *regular* if it is semiregular and closed under relativization — the operation of replacing the universe by an $\mathcal{L}$ definable set. Let us give a concrete definition for the case of graphs:

Definition 5.

1. A language of graphs $\mathcal{L}$ is said to be *semiregular* if:
    - All atomic formulas are in $\mathcal{L}$.
    - If $\varphi \in \mathcal{L}$ then $\neg\varphi \in \mathcal{L}$.

modified:2015-10-21

986 revision:2015-10-21_2

- If $\varphi, \psi \in \mathcal{L}$ then $\varphi \wedge \psi \in \mathcal{L}$.
- If $\varphi(x) \in \mathcal{L}$ then $\exists x.\, \varphi(x) \in \mathcal{L}$.
- (Weak substitution) If $\varphi(x, y), \psi \in \mathcal{L}$ and φ is anti-reflexive and symmetric, then there exists a sentence $\psi' \in \mathcal{L}$ such that

  $$G \models \psi \iff (V, \{(x, y) \in V \times V \mid G \models \varphi(x, y)\}) \models \psi'$$

  where $G = (V, E)$ is a graph.

2. $\mathcal{L}$ is said to be *regular* if in addition

   - (Full substitution) Let $\varphi(x)$ and ψ be formulas in $\mathcal{L}$ and denote $V_\varphi = \{x \in V \mid G \models \varphi(x)\}$. Then there exists a sentence $\psi' \in \mathcal{L}$ such that

     $$G \models \psi \iff (V_\varphi, G[V_\varphi]) \models \psi'$$

     where $G = (V, E)$ is a graph, and $G[V']$ is the graph spanned on the vertex subset $V' \subset V$.

In order to get the minimal semiregular language that can express a property K we use generalized quantifiers or *Lindström quantifiers*. Let K be a property of graphs. We think of K as the set of all graphs having this property. As a graph property, K is closed under isomorphism. Given K we define the graph language $\mathcal{L}(Q_K)$ as follows.

Definition 6.

1. Let K be a graph property. The set of formulas $\mathcal{L}(Q_K)$ is the closure of the atomic formulas by the conjunction and negation connectives, the existential quantifier and another quantifier Q_K. The syntax of the new quantifier is $Q_K xy.\varphi(x, y, \bar{a})$ where $\varphi(x, y, \bar{a}) \in \mathcal{L}(Q_K)$ is an antireflexive and symmetric formula in which x, y are free variables and $\bar{a}$ are parameters.

2. Given a graph $G = (V, E)$, the satisfaction of formulas of the form $Q_K xy.\, \varphi(x, y, \bar{a})$ is determined by

   $$G \models Q_K xy.\, \varphi(x, y, \bar{a}) \iff (V, \{(x, y) \in V \times V \mid G \models \varphi(x, y, \bar{a})\}) \in K.$$

modified:2015-10-21
986 revision:2015-10-21_2

Of course, Lindström quantifiers are not restricted to graphs and can be defined for any vocabulary (indeed, usually that is the case).

Here are three simple examples over sets: First notice that

$$\forall x.\, \varphi(x) \Longleftrightarrow Q_{\{U\}}x.\, \varphi(x)$$

where U is the universe set. The expression on the right hand side means that we consider the set of all x's for which $\varphi(x)$ holds, and then we check if this set belongs to the singleton $\{U\}$, that is, if it is the whole universe. Similarly $\exists x.\, \varphi(x) \Longleftrightarrow Q_{P(U)\setminus\{\varnothing\}}x.\, \varphi(x)$ where $P(U)$ is the power set of U. As a final example we mention that the parity quantifier of [23] mentioned above may be expressed as a Lindström quantifier e. g. by writing $\oplus x.\, \varphi(x) \Longleftrightarrow Q_{\{A\in P(U) \mid |A| \text{ is even}\}}x.\, \varphi(x)$.

The basic result regarding the Lindström quantifiers is that for any property K, the language $\mathcal{L}(Q_K)$ is the smallest semiregular language that can express K [26].

The Lindström quantifier Q_K acts as an oracle that let us check if the graph that we get by replacing the edge relation with a defined one is in K. We can create more expressive languages by allowing more freedom in defining the graphs to be queried. We describe three variants here, in increasing expressive power.

Given a graph property K we define the *relativized* language $\mathcal{L}^{\mathrm{rl}}(Q_k)$ similarly to the above, but this time we have another formula defining which of the original vertices are included in the queried graph. We denote the resulting language by $\mathcal{L}^{\mathrm{rl}}(Q_k)$.

Definition 7.

1. Let K be a graph property. The set of formulas $\mathcal{L}^{\mathrm{rl}}(Q_k)$ is the closure of the atomic formulas by first order operations together with additional quantifier Q_K. The syntax of the new quantifier is given by $Q_K vxy.\, \psi(v,\bar{a}), \varphi(x,y,\bar{a})$ where $\varphi(x,y,\bar{a}) \in \mathcal{L}^{\mathrm{rl}}(Q_k)$ is an antireflexive and symmetric formula in which x, y are free variables, $\psi(v,\bar{a}) \in \mathcal{L}^{\mathrm{rl}}(Q_k)$ is a formula in which v is free and $\bar{a}$ are parameters.

2. Given a graph $G = (V, E)$, the satisfaction of formulas of the form $Q_K vxy.\, \psi(v,\bar{a}), \varphi(x,y,\bar{a})$ is determined as follows. Let $V|_\psi = \{v \in V \mid \psi(v,\bar{a})\}$. Then

$$\begin{aligned} &G \models Q_K vxy.\, \psi(v,\bar{a}), \varphi(x,y,\bar{a}) \Longleftrightarrow \\ &(V|_\psi, \{(x,y) \in V|_\psi \times V|_\psi \mid \varphi(x,y,\bar{a})\}) \in K. \end{aligned}$$

modified:2015-10-21
986 revision:2015-10-21_2

It is easy to verify that $\mathcal{L}^{\mathrm{rl}}(Q_k)$ is regular for any K, Moreover, it is the inclusion minimal regular language that is able to express K.

In the next variant we allow to redefine the other predicate in the graph vocabulary as well, namely, the equality predicate. This means that the vertex set will be the quotient set of some equivalence relation. We denote the resulting language by $\mathcal{L}^{\mathrm{eq}}(Q_k)$.

Definition 8.

1. Let K be a graph property. The set of formulas $\mathcal{L}^{\mathrm{eq}}(Q_k)$ is the closure of the atomic formulas by first order operations together with additional quantifier Q_K. The syntax of the new quantifier is given by $Q_K vuwxy.\psi(v,\bar{a}), \varphi_=(u,w,\bar{a}), \varphi_\sim(x,y,\bar{a})$ where $\varphi_\sim(x,y,\bar{a}) \in \mathcal{L}^{\mathrm{eq}}(Q_k)$ is an antireflexive and symmetric formula in which x, y are free variables, $\varphi_=(u,w,\bar{a}) \in \mathcal{L}^{\mathrm{eq}}(Q_k)$ is a reflexive, symmetric and transitive formula in which u, w are free variables, $\psi(v,\bar{a}) \in \mathcal{L}^{\mathrm{eq}}(Q_k)$ is a formula in which v is a free variable and $\bar{a}$ are parameters.

2. Given a graph $G = (V,E)$, the satisfaction of formulas of the form $Q_K vuwxy.\psi(v,\bar{a}), \varphi_=(u,w,\bar{a}), \varphi_\sim(x,y,\bar{a})$ is determined as follows. Let $V' = \{v \in V \mid \psi(v,\bar{a})\}$ and let R be the equivalence relation induced over V' by $\varphi_=$ (that is, $R = \{(u,w) \in V' \times V' \mid \psi_=(u,w,\bar{a})\}$). Then the semantics of Q_K in $\mathcal{L}^{\mathrm{eq}}(Q_K)$ is given by
$$G \models Q_K vuwxy.\psi(v,\bar{a}), \varphi_=(u,w,\bar{a}), \varphi_\sim(x,y,\bar{a}) \iff$$
$$(V'/R, \{([x],[y]) \in (V'/R) \times (V'/R) \mid \varphi_\sim(x,y,\bar{a})\}) \in K.$$

Generally, care must be taken in order to make sure that the edges induced by $\psi_\sim$ are well defined.

The strongest variant lets us redefine the vertices as *tuples* of vertices. Let l be an integer. The syntax and the semantics of $\mathcal{L}^{\mathrm{tu}}(Q_k)$ are simply vectorized versions of the syntax and semantics of $\mathcal{L}^{\mathrm{eq}}Q_k)$. The formula $\psi(\bar{v},\bar{a})$ determines which l-tuple is a vertex, $\varphi_=(\bar{u},\bar{w},\bar{a})$ is an equivalence relation defining equality between l-tuples and $\varphi_\sim(\bar{x},\bar{y},\bar{a})$ defines the edge set of the graph.

Definition 9.

1. Let K be a graph property. The set of formulas $\mathcal{L}^{\mathrm{tu}}(Q_k)$ is the closure of the atomic formulas by first order operations together with additional quantifier Q_K. The syntax of the new quantifier is given by $Q_K \bar{v}\bar{u}\bar{w}\bar{x}\bar{y}.\psi(\bar{v},\bar{a}), \varphi_=(\bar{u},\bar{w},\bar{a}), \varphi_\sim(\bar{x},\bar{y},\bar{a})$ where:

modified:2015-10-21

986 revision:2015-10-21_2

(a) all the vectors $\bar{v}, \bar{u}, \bar{w}, \bar{x}$ and $\bar{y}$ are of the same length denoted henceforth by l;

(b) $\varphi_{\sim}(\bar{x}, \bar{y}, \bar{a}) \in \mathcal{L}^{\mathrm{tu}}(Q_k)$ is an antireflexive and symmetric formula in which $\bar{x}, \bar{y}$ are l-tuples of free variables;

(c) $\varphi_{=}(\bar{u}, \bar{w}, \bar{a}) \in \mathcal{L}^{\mathrm{tu}}(Q_k)$ is a reflexive, symmetric and transitive formula in which $\bar{u}, \bar{w}$ are l-tuples of free variables;

(d) $\psi(\bar{v}, \bar{a}) \in \mathcal{L}^{\mathrm{tu}}(Q_k)$ is a formula in which $\bar{v}$ is an l-tuple of free variables; and

(e) $\bar{a}$ are parameters.

2. Given a graph $G = (V, E)$, the satisfaction of formulas of the form $Q_K \bar{v}\bar{u}\bar{w}\bar{x}\bar{y}.\psi(\bar{v}, \bar{a}), \varphi_{=}(\bar{u}, \bar{w}, \bar{a}), \varphi_{\sim}(\bar{x}, \bar{y}, \bar{a})$ is determined as follows. Let $V' = \{\bar{v} \in V^l \mid \psi(\bar{v}, \bar{a})\}$ and let R be the equivalence relation induced over V' by $\varphi_{=}$. Then the semantics of Q_K in $\mathcal{L}^{\mathrm{tu}}(Q_K)$ is given by

$$G \models Q_K \bar{v}\bar{u}\bar{w}\bar{x}\bar{y}.\psi(\bar{v}, \bar{a}), \varphi_{=}(\bar{u}, \bar{w}, \bar{a}), \varphi_{\sim}(\bar{x}, \bar{y}, \bar{a}) \Longleftrightarrow$$
$$(V'/R, \{([\bar{x}], [\bar{y}]) \in (V'/R) \times (V'/R) \mid \varphi_{\sim}(\bar{x}, \bar{y}, \bar{a})\}) \in K.$$

As before, one must verify that the edges induced by $\psi_{\sim}$ are well defined.

1.4 Results

Our results are of mixed nature. We show that there are regular languages able to express any first order sentence, connectivity and k-colorability for any fixed k, and still obey the zero-one law for $G(n, p)$ for any constant $0 < p < 1$. On the other hand, for the same model of random graphs we show that in any semiregular language able to express Hamiltonicity, there is a sentence with no limiting probability.

Our results are concerned with connectivity, k-colorability and Hamiltonicity, thus we define the following sets of graphs:

1. Conn, the set of all connected graphs,

2. Ham, the set of all Hamiltonian graphs, that is, the set of graphs having a Hamilton cycle, and

3. Ch_k, the set of all graphs having chromatic number k.

modified:2015-10-21
986 revision:2015-10-21_2

Now we can state our results.

Theorem 10. *For every constant* $0 < p < 1$ *and for every sentence* φ *in* $\mathcal{L}^{\mathrm{tu}}(Q_{\textsc{Conn}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots)$,

$$\lim \Pr[G(n,p) \models \varphi] \in \{0,1\}.$$

Remark 11. Let $\textsc{Pln}$ be the set of all planar graphs. Using the same technique used for proving Theorem 10 one may add a planarity Lindström quantifier to the language appearing in the statement of the theorem. The following result will be published elsewhere:

For every constant $0 < p < 1$ and for every sentence φ in the graph language $\mathcal{L}^{\mathrm{tu}}(Q_{\textsc{Conn}}, Q_{\textsc{Pln}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots)$,

$$\lim \Pr[G(n,p) \models \varphi] \in \{0,1\}.$$

Theorem 12. *For any constant* $0 < p < 1$, *the pair* $(\mathcal{L}(Q_{\textsc{Ham}}), G(n,p))$ *can interpret arithmetic for some function* $f = \Omega(\log\log\log n)$.

Remark 13.

1. The interpretation mentioned in Theorem 12 is theoretically explicitly given.

2. No effort was made on getting the best behavior for f, and clearly $f = \Omega(\log\log\log n)$ is far from optimal. In fact, with some effort one can show that the pair $(\mathcal{L}(Q_{\textsc{Ham}}), G(n,p))$ can interpret arithmetic for some *linear* function.

3. Having arithmetization means that $\mathcal{L}(Q_{\textsc{Ham}})$ violates even the modular limit law of Kolaitis and Kopparty.

The following corollary is an immediate consequence of Theorem 12 and the fact that $\mathcal{L}(Q_{\textsc{Ham}})$ is the inclusion minimal semiregular language in which Hamiltonicity is expressible.

Corollary 14. *Let* $\mathcal{L}$ *be a language that can express Hamiltonicity and let* $0 < p < 1$ *be constant, then the pair* $(\mathcal{L}, G(n,p))$ *has arithmetization.*

This answers the question of Blass and Harary [5].

986 revision:2015-10-21_2 modified:2015-10-21

1.5 Discussion

This study suggests a notion of simplicity for properties of graphs. Properties for which the first order closure obeys the zero-one law are simpler than properties P for which $\mathcal{L}(Q_P)$ has nonconverging sentences, or worse, are able to interpret arithmetic. There is an intermediate level of properties for which $\mathcal{L}(Q_P)$ does not obey the zero-one law but every sentence has a limiting probability. Studying this question for other properties of graphs seems natural and interesting. Obtaining general criteria for graph properties to obey or fail the zero-one law seems within grasp.

The same question arises also for other situations in which the zero-one law holds. In particular we are interest in the behavior of the same languages when the random graph is $G(n, p = n^{-\alpha})$ and $0 < \alpha < 1$ is irrational. As mentioned above, $\mathcal{FO}$ obeys the zero-one law in this case [29]. The partial results that we have [18] suggests that at least $\mathcal{L}(Q_{\text{HAM}})$ behave differently at this situation, but this is still a work in progress.

Looking at the proof of Theorem 10 one sees that while the language has strictly stronger expressive power, it defines the same family of definable sets as $\mathcal{FO}$. We were also interested in the question of finding a language with more definable sets than $\mathcal{FO}$, that still obeys the zero-one laws. Lately a rather natural language having these properties was found ([19], paper in preparation).

1.6 Notation

Along the paper we use standard graph theory notions and notation.

Let $G = (V, E)$ be a graph. We denote the neighborhood of a vertex v by $N(v) = \{u \in V | u \sim v\}$. We denote the degree of v by $d(v) = |N(v)|$. Given a vertex set $S \subset V$ we denote the neighborhood of v in S by $N(v, S) = N(v) \cap S$ and the degree of v in S by $d(v, S) = |N(v, S)|$. Given two vertices u, V their *codegree* is the number of common vertices, and is denoted here by $\text{codeg}(u, v)$. We refer the reader to [32], [8] or [10] for general graph theory monographs.

Let $\mathcal{L}$ be a formal language for the class of graphs and let $G = (V, E)$ be a graph. A vertex set $A \subset V$ is said to be *definable in* $\mathcal{L}$ (also $\mathcal{L}$-*definable*), if there is a formula $\varphi(x) \in \mathcal{L}$ with x being a free variable such that $A = \{a \mid G \models \varphi(a)\}$.

We use the standard "Big O" asymptotic notation of Bachmann and

modified:2015-10-21
986 revision:2015-10-21_2

986 revision:2015-10-21_2 modified:2015-10-21

Landau. Let $f(n), g(n)$ be two positive functions whose domain is $\omega = \{0, 1, 2, \dots\}$. We say that $f = O(g)$ if there is a constant C such that $f(n) \leq Cg(n)$ for every integer n. We say that $f = \Omega(g)$ if $g = O(f)$, and that $f = \Theta(g)$ if both $f = O(g)$ and $f = \Omega(g)$. If $f/g \to 0$ as $n \to \infty$ we write $f = o(g)$, if $f = o(g)$ then we also write $g = \omega(f)$. In particular, we may use $f = o(1)$ and $f = \omega(1)$ to denote functions tending to zero and to infinity respectively. We also write $f = (1 \pm \epsilon)g$ if there exists a constant N such that $(1-\epsilon)g(n) \leq f(n) \leq (1+\epsilon)g(n)$ for every $n > N$. We add a diacritical tilde above the relation symbol to denote asymptotic relation which holds up to polylogarithmic factors. That is, $f = \tilde{O}(G)$ means that $f = O(g \cdot h)$ where h is some polylogarithmic function. In the case of the strict relations the tilde means "by more than a polylogarithmic function". For example, $f = \tilde{\omega}(g)$ means that $f/gh \to \infty$ for *any* polylogarithmic h.

We use ln for base $e = 2.718281828\dots$ logarithms and log for base two logarithms.

2 Connectivity and chromatic number

In this section we prove Theorem 10.

Proof of Theorem 10. Let p be a constant, $0 < p < 1$. Assume first that φ is a sentence in the language $\mathcal{L} = \mathcal{L}(Q_{\text{CONN}}, Q_{\text{CH}_2}, Q_{\text{CH}_3}, \dots)$, which is the first-order closure of the quantifiers $Q_{\text{CONN}}, Q_{\text{CH}_2}, Q_{\text{CH}_3}, \dots$ (using the weakest sense of a Lindström quantifier as in Definition 6). We wish to show that the limiting probability of φ in $G(n,p)$ is either zero or one. We use induction on the structure of φ.

Claim 10.1: Let $\varphi(\bar{a})$ be a formula in $\mathcal{L}$, where $\bar{a}$ are free variables (seen as parameters of φ). Then the limiting probability of $\varphi(\bar{a})$ in $G(n,p)$ with $0 < p < 1$ is either zero or one, depending only on the relations among $\bar{a}$.

Proof of Claim 10.1: By induction on the structure of the formula. For an atomic formula it is clear (that is, the truth value of $a_1 = a_2$ and $a_1 \sim a_2$ depends only on the relations among a_1 and a_2...).

If φ is of the form $\neg\psi$ or $\psi_1 \wedge \psi_2$ then the claim is immediate.

Assume that $\varphi(\bar{a})$ is of the form $\varphi(\bar{a}) = Q\bar{y}.\psi(\bar{a}, \bar{y})$ where Q is one of $\exists, Q_{\text{CONN}}, Q_{\text{CH}_k}$ and $\bar{y}$ is of length one or two. Consider the equivalence relation $E_{\bar{a}}$ over vertices defined by

$$xE_{\bar{a}}y \iff \forall a \in \bar{a}.\ [(x = a \leftrightarrow y = a) \wedge (x \sim a \leftrightarrow y \sim a)],$$

and let $B_i, 1 \leq i \leq k_* = |\bar{a}| + 2^{|\bar{a}|}$ be the equivalence classes in $[n]/E_{\bar{a}}$. By the induction hypothesis, the limiting probability of $\psi(\bar{a}, \bar{y})$ depends only on the relations among $\bar{a}$ and $\bar{y}$. In other words, if $Q = \exists$ then the aforementioned limiting probability is determined by the relations among $\bar{a}$ and the i for which $b \in B_i$. If $Q = Q_{\text{CONN}}$ or $Q = Q_{\text{CH}_k}$ then the limiting probability depends on relations among $\bar{a}$, the equivalence classes containing the two variables of $\bar{y}$ and the relations between these variables.

Now, for the first time, probability enters the proof. Each of the singletons $\{a_i\}$ is an equivalence class. Let B_i be an equivalence class defined by some adjacency pattern between a vertex and $\bar{a}$, and let $t = t(i)$ be the number of adjacencies in the pattern. The probability that a given vertex belongs to B_i is then $p_* = p_*(i) := p^t(1-p)^{|\bar{a}|-t}$. Clearly and importantly, the event "v belongs to B_i" is independent of all other events of the form "u belongs to B_j". Hence the size of the equivalence class follows a binomial distribution with parameters $n - |\bar{a}|$ and p_*, and in particular a.a.s. all equivalence classes are of size which is tightly concentrated around its linear sized $(p_*(n - |\bar{a}|))$ mean.

Assume first that Q is the existential quantifier. If for any of the (finitely many) B_i's the limiting probability of $\psi(\bar{a}, y)$ is one for $y \in B_i$, then φ also has limiting probability one. Otherwise φ's limiting probability is zero.

Next assume that $\varphi(\bar{a})$ is of the form $\varphi(\bar{a}) = Q_{\text{CONN}} y_1, y_2.\psi(\bar{a}, y_1, y_2)$. We want to show that the limiting probability for the connectivity of the graph $F = ([n], \{\{y_1, y_2\} \mid \psi(\bar{a}, y_1, y_2)\})$ depends solely on the relations between the members of $\bar{a}$ and themselves. If $\bar{a}$ is the empty sequence, then F is either the empty graph over $[n]$ (which is disconnected), the complete graph over $[n]$ (connected), the original graph or its compliment. In the last two cases, since the original graph is sampled from $G(n,p)$, it is a.a.s. connected (this is a well known fact. See, e.g., [21]). All in all, the limiting probability is either zero or one as required.

Assume now that $\bar{a}$ is not empty. Notice that by the induction hypothesis, once $\bar{a}$ and the relations among its members are given, the limiting probability of $\psi(\bar{a}, y_1, y_2)$ depends only upon the classes of $[n]/E_{\bar{a}}$ containing y_1 and y_2, whether $y_1 \sim_{G(n,p)} y_2$ and whether $y_1 = y_2$. In particular, if both y_1 and y_2 are not singleton, then we are in one of four situations. It may be that for every $y_1 \in B_1, y_2 \in B_2$ the formula $\psi(\bar{a}, y_1, y_2)$ is true (for given $\bar{a}$), or that for every such pair the formula $\psi(\bar{a}, y_1, y_2)$ is false, or that the truth value of $\psi(\bar{a}, y_1, y_2)$ depends *solely* on the truth value of $y_1 \sim_{G(n,p)} y_2$ (the last case counts as two, either $\psi(\bar{a}, y_1, y_2)$ agrees with $y_1 \sim_{G(n,p)} y_2$ or they disagree).

986 revision:2015-10-21_2 modified:2015-10-21

Consider the graph H whose vertex set is the quotient set $[n]/E_{\bar{a}}$ and its edge set is determined by

$$B_1 \sim_H B_2 \Longleftrightarrow \exists y_1 \in B_1, y_2 \in B_2.\, \psi(\bar{a}, y_1, y_2).$$

Notice that we allow loops in H. We argue that if H is disconnected then so is F, and if H is connected then F is connected with probability tending to one with n. Let u and w be two vertices of F and assume that there is a path $v_0 = u, v_1, \ldots, v_l = w$ in F connecting u and w. Then $[v_0], [v_1], \ldots, [v_l]$ is a walk connecting u and w in H (formally this may not be a walk since it may be that $[v_i] = [v_j]$ without a loop. Still it contains a path connecting $[v]$ and $[w]$). Hence, if H is disconnected then F is disconnected as well.

Assume now that H is connected and let u, w be two vertices in F. For the moment, assume also that $[u] \neq [w]$. Denote the vertices of the shortest path connecting $[u]$ and $[w]$ in H by $[v_0 = u], [v_1], \ldots, [v_l = w]$. We find a path connecting u and w in F by induction on the length of the path as follows. Start by denoting u as v'_0 and assume that we have a path in F connecting u to v'_i where $v'_j \in [v_j]$ for every $0 \leq j \leq i$. Since $[v_i] \sim_H [v_{i+1}]$, we know that there are two vertices $y_1 \in [v_i]$ and $y_2 \in [v_{i+1}]$ such that $\psi(\bar{a}, y_1, y_2)$ holds. If either $[v_i]$ or $[v_{i+1}]$ is a singleton (or both), then, by definition of H's vertices, all the vertices of $[v_i]$ relate in the same manner to all the vertices of $[v_{i+1}]$. Hence we can take $v'_{i+1} = v_{i+1}$. If both $[v_i]$ and $[v_{i+1}]$ are not singletons, then, as argues above, both are of size linear in n. If the value of $\psi(\bar{a}, y_1, y_2)$ does not depend on $y_1 \sim_{G(n,p)} y_2$, we can simply pick $v'_{i+1} = v_{i+1}$. Otherwise we use the fact that the probability of v'_i not having a neighbor or a non-neighbor in a vertex set of size cn is exactly $p^{cn} + (1-p)^{cn} \leq c'^n$ where $0 < c' < 1$ is some constant. Hence, with probability tending to one exponentially fast we can find a vertex $v'_{i+1} \in [v_{i+1}]$ satisfying $\psi(\bar{a}, v'_i, v'_{i+1})$. When picking v'_{l-1} we need to take extra care and make sure that it is a neighbor of w in F as well. This is easily done using a similar consideration (and similar computations).

Assume now that $[u] = [w]$. In this case we consider a closed walk instead of the shortest path. That is, we denote $[v_0 = u], [v_1], [v_2 = w]$, where $[v_1] \neq [u]$. Such $[v_1]$ exists since we are in the case of nonempty $\bar{a}$, meaning that the number of vertices in the connected graph H is greater than one. The rest of the proof is identical to the above.

We have shown that with probability tending exponentially fast to one, the connectivity of F depends only on the connectivity of H. By the definition of H, it is connected depending only on $\bar{a}$.

modified:2015-10-21
986 revision:2015-10-21_2

Finally let $Q = \textsc{Ch}_k$ for some constant $k \in \mathbb{N}$. We look again on the graphs F and H defined above. By arguments similar to the above, there are k_* vertices $v_1, \dots, v_{k_*} \in F$ such that $v_i \in B_i$ and $v_i \sim_F v_j \Leftrightarrow B_i \sim_H B_j$. Hence F has a copy of H as a subgraph and therefore $\chi(F) \geq \chi(H)$. We argue that with probability tending to one there are two possibilities: either $\chi(F) = \chi(H)$ or $\chi(F) = \Omega(n/\ln n)$. If for every i the vertices of B_i form an independent set in F, then we can color F by assigning the color of B_i to all its vertices. Hence in this case $\chi(F) = \chi(H)$. Assume now that there is a set B_i that is not an independent set and let $n' = |B_i|$ and recall that a.a.s. $n' = cn(1 + o(1)))$. The graph spanned by F on B_i is either a clique, a spanned subgraph of $G(n, p)$ or the complement of a spanned subgraph of $G(n, p)$. In the first case the chromatic number of the spanned graph is n', so $\chi(F) = \omega(n/\log n)$. In the second and third cases the chromatic number of $F[B_i]$ is in fact the chromatic number of a random graph with n' vertices and edge probability p' being equal to p or to $1 - p$. In an exciting paper [6] Bollobás showed that with probability tending to one this number is $(1/2+o(1))n'/\log_b n'$ where $b = 1/(1-p')$. Thus in this case the chromatic number of F is of order $n/\ln n$.

Summarizing the above argument we get that the chromatic number of $([n], \{\{y_1, y_2\} \mid \psi(\bar{a}, y_1, y_2)\})$ depends only on the relations among $\bar{a}$ and it is either a specific constant or growing to infinity with n. Therefore the limiting probability of $Q_{\textsc{Ch}_k} xy.\, \psi(\bar{a}, x, y)$ is either zero or one, depending only on the relations among $\bar{a}$.

End of proof of Claim 10.1.

The theorem for the case of formulas in $\mathcal{L}$ follows immediately as φ is a sentence in $\mathcal{L}$ with no free variables.

We argue now that the proof works, mutatis mutandis, also for $\varphi \in \mathcal{L}^{\text{tu}} = \mathcal{L}^{\text{tu}}(Q_{\textsc{Conn}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots)$, where $\mathcal{L}^{\text{tu}}(Q_{\textsc{Conn}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots)$ is the first order closure of $Q_{\textsc{Conn}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots$ with the syntax and semantics as in Definition 9. The proof still works by quantifier elimination, and the key observation is that we can still partition the vertices into equivalence classes that are either singletons or sets of polynomial size. We restrict ourselves to a description of the changes in the induction step in Claim 10.1. Assume that φ is of the form $\varphi = Q_K \bar{v}\bar{u}\bar{w}\bar{x}\bar{y}.\psi(\bar{v}, \bar{a}), \varphi_{=}(\bar{u}, \bar{w}, \bar{a}), \varphi_{\sim}(\bar{x}, \bar{y}, \bar{a})$, where k is one of $Q_{\textsc{Conn}}, Q_{\textsc{Ch}_2}, Q_{\textsc{Ch}_3}, \dots$, and $\psi, \varphi_{=}$ and $\varphi_{\sim}$ are as in Definition 9. Let $l = |\bar{v}|$, and notice the number l-tuples of vertices is n^l. The relations between the l-tuples and the elements of $\bar{a}$ will divide the set of l-tuples into $(|\bar{a}| + 2^{|\bar{a}|})^l$ — finitely many — equivalence classes, denoted as above by

modified:2015-10-21
986 revision:2015-10-21_2

B_i. The size of an equivalence class defined by s equalities to elements of $\bar{a}$ (and thus $|\bar{a}| - s$ inequalities — the adjacencies does not matter here) is $\Theta(n^{l-s})$. To see this, consider an equivalence class B_i defined by s equalities to elements of $\bar{a}$. There are $(n - |\bar{a}|)^{l-s}$ tuples[4] that satisfy the s equalities. Call the variables that are not required to be equal to any of the elements of $\bar{a}$ an inequality variable. Then the number of l-tuples that additionally satisfy all the adjacency relations for the first inequality variable is a Binomial random variable with parameters $(n - |\bar{a}|)^{l-s}$ and p^*, where p^* is a constant of the form $p^r(1-p)^t$ (r and t are two constants satisfying $0 < r, t < |\bar{a}|$). Hence a.a.s. the number of such l-tuples is of order n^{l-s}. Applying the same argument to the rest of the (finitely many) inequality variables, gives that B_i is of size $\Theta(n^{l-s})$.

Now, ψ defines the set of vertices, and it will be a union of equivalence classes, hence it is either of constant size or of polynomial size in n. The affect of $\varphi_=$ is similar. Indeed, $\varphi_=(\bar{x}, \bar{y})$ depends on the classes B_i, B_j that contain $\bar{x}$ and $\bar{y}$, but also on the relations between $\bar{x}$ and $\bar{y}$. Still, there are only finitely many options for such relations, and an argument identical to the argument above gives that the number of l-tuples in each vertex of the defined graph remains either constant or polynomial in n. The rest of the proof remains very similar. $\square$

3 Hamiltonicity

In this section we prove Theorem 12 by showing that using properties of $G(n,p)$ we can encode any *second order* sentence into a sentence of $\mathcal{L}(Q_{\text{HAM}})$ such that the sentences are equivalent on a set of vertices with size tending to infinity at a controlled pace as n grows.

Let us start with a brief overview of the proof. Using the Q_{HAM} quantifier we shall be able to express $|A| \leq |B|$ where A and B are (definable) vertex sets. Next we would like to find sets that are: definable, small enough so that all their subsets are definable, and finally also large enough — tend to infinity as n grows in an appropriate rate. Given the ability to express equality of definable sets it seems plausible to look on sets of the form "all vertices having degree d". Still, it is not obvious how to continue, as the set of all vertices of degree np is too large, while the set of vertices of, say, minimal degree is too small. We shall define such a set having size $\Theta(\log \log n)$. This enables us to

[4] If some of the elements of $\bar{a}$ are identical, we may have a few more candidate tuples.

modified:2015-10-21
986 revision:2015-10-21_2

interpret *monadic* second order on this set, which almost suffice considering the fact that the induced graph on this set should be quite random and then an application of [22] should give arithmetization. We shall prefer to give a direct proof that will require less randomness from the induced graph, and it is therefore technically less involved.

We begin by defining two graph properties that are needed later for the encoding scheme.

Definition 15. Let $\mathcal{L}$ be a language and let $\varphi(x, \bar{y}) \in \mathcal{L}$ be a formula in which x is a free variable. We say that a given graph $G = (V, E)$ *allows φ-powerset representation*, if there exists a vertex sequence $\bar{b} \in V^{|\bar{y}|}$ such that for every subset $A \subset S := \{x \in V \mid G \models \varphi(x, \bar{b})\}$ there is a vertex $v \in V$ such that $N(v, S) = A$.

If G allows φ-powerset representation then we can represent monadic second order variables on S by vertices from V. Generally, this is useful if S is large enough. In order to encode second order sentences we need a bit more:

Definition 16. Let $\mathcal{L}$ be a language of graphs. We say that a graph $G = (V, E)$ *allows $(\varphi_0(x, \bar{y}_0), \varphi_1(x, \bar{y}_1), f)$-double powerset representation* when

1. $f \colon \mathbb{N} \to \mathbb{N}$ is a monotone increasing function.
2. There exists a vertex sequence $\bar{b}_0 \in V^{|\bar{y}|}$ such that for every subset $A \subset S_{\bar{b}_0} := \{x \in V \mid G \models \varphi(x, \bar{b}_0)\}$ there is a vertex $v \in V$ such that $N(v, S) = A$. That is, G allows φ_0-powerset representation.
3. Let $S = S_{\bar{b}_0}$ for $\bar{b}_0$ as in Item 2 above. Then
   (a) There is a finite vertex sequence $\bar{b}_1 \in V^{|\bar{y}_1|}$ such that the set $B_{\bar{b}_1} := \{x \in S \mid G \models \varphi_1(x, \bar{b}_1)\}$ has at least $f(|V|)$ members.
   (b) For every $A \subseteq B_{\bar{b}_1}$ there is a vertex $v \in S$ such that $N(v, S) = A$.

Notice that in the definition above every subset of $B := B_{\bar{b}_1}$ is represented as a neighborhood in B of a vertex from S. The set B will be the set over which we shall be able to encode second order sentences. Hence we are interested in a lower bound for its size, which is the role of f.

The following lemma demonstrates the usefulness of the definition above — basically saying that when a graph G on n vertices allows $(\varphi_0, \varphi_1, f)$-double powerset representation, then we can encode arithmetic over the set $\{1, \ldots, f(n)\}$. This lemma uses ideas similar to the main ideas of [22].

modified:2015-10-21
986 revision:2015-10-21_2

Lemma 17. *Let $\mathcal{L}$ be a semiregular language and let $G = (V, E)$ be a graph. Assume that G allows $(\varphi_0, \varphi_1, f)$-double powerset representation. Then there is a set B of size at least $f(|V|)$ such that for every second order sentence φ there is a sentence $\varphi' \in \mathcal{L}$ satisfying*

$$G[B] \models \varphi \Longleftrightarrow G \models \varphi'.$$

Proof. Let $G = (V, E)$ be a graph that allows $(\varphi_0(x, \bar{y}_0), \varphi_1(x, \bar{y}_1), f)$-double powerset representation and let $\bar{b}_0$ and $\bar{b}_1$ be the vertex sequences for which Definition 16 is realized. Denote by S and B the sets $S_{\bar{b}_0}$ and $B_{\bar{b}_1}$ (respectively) appearing in the definition. We start by defining a linear order over the vertices of B. Fix some linear order on B and mark the vertices of B according to that order by $1, 2, \ldots, b$. Since G allows $(\varphi_0, \varphi_1, f)$-double powerset representation, for every $1 \leq i \leq b$ there is a vertex v_i in S such that $N(v_i, B) = \{1, \ldots, i\}$. Also, again by G allowing a $(\varphi_0, \varphi_1, f)$-double powerset representation, there is a vertex $v_<$ such that $N(v_<, S) = \{v_1, \ldots, v_b\}$. We shall use vertices like $v_<$ to represent a linear order over B.

We want a vertex $v_< \in V$ satisfying the formula $\mathrm{Order}(v_<)$ defined and explained immediately:

$$\begin{aligned}\mathrm{Order}(v_<) :=& [\forall s, t \in S. \, (s \sim v_< \wedge t \sim v_<) \rightarrow \\ & ((N(s, B) \subsetneq N(t, B)) \vee (N(t, B) \subsetneq N(s, B)))] \wedge \\ & \forall b \in B. \, \exists s, t \in N. \, [(v_<). N(t, B) \triangle N(s, B) = \{b\}].\end{aligned}$$

That is, for $\mathrm{Order}(v_<)$ to hold for some vertex $v_< \in V$ the following must happen:

1. Ordering the neighbors of $v_<$ in S according to inclusion of their neighborhoods in B gives a linear order.

2. The number of elements in that order $(d(v_<, S))$ is at least $|B|$ as any element of B separates between the neighborhoods of a pair of $v_<$ neighbors. (Also $d(v_<, S) \leq |B| + 1$).

Clearly $v_<$ induces a linear ordering over the elements of B. Let us give a concrete definition. Assume that $v_< \in V$ satisfies $\mathrm{Order}(v_<)$. We define

$$b_1 <_{v_<} b_2 := \forall s \in S. \, [((s \sim v_<) \wedge (s \sim b_2)) \rightarrow (s \sim b_1)].$$

modified:2015-10-21
986 revision:2015-10-21_2

It is not difficult to observe that $<_{v_<}$ is a linear order over B. Again, since G allows $(\varphi_0, \varphi_1, f)$-double powerset representation we can find such $v_<$ for every linear order $<$ over B.

Having order we can describe the encoding of second order sentences. Let φ be a second order sentence. We want to write a sentence $\varphi' \in \mathcal{L}$ such that $G[B] \models \varphi \Leftrightarrow G \models \varphi'$. Monadic variables can be replaced using vertices from S by replacing, e.g. $\exists A^1.\, \psi(A)$ by $\exists v_A \in S.\, \psi'(v_A)$ where ψ' is the sentence we get by replacing every occurrence of $x \in A$ in ψ by $x \sim v_A$.

For binary variables we need a bit more. Since G allows $(\varphi_0, \varphi_1, f)$-double powerset representation, for every pair of vertices $x, y \in B$ we have a vertex $s_{\{x,y\}} \in S$ such that $N(s_{\{x,y\}}, B) = \{x, y\}$. Every binary relation R can be represented by two sets of pairs — the set of all the pairs x, y such that $x \leq_{v_<} y$ and xRy and the set of all the pairs x, y such that $x >_{v_<} y$ and xRy. We represent binary relations using this idea. That is, we replace every occurrence of the form $\exists R^2.\, \psi(R)$ appearing in φ by $\exists v_{R,\{1,2\}}, v_{R,\{2,1\}}.\, \psi'(v_{R,\{1,2\}}, v_{R,\{2,1\}})$. In order to get $\psi'(v_{R,\{1,2\}}, v_{R,\{2,1\}})$ from $\psi(R)$ we replace occurrences of the form xRy in ψ by

$$((x \leq_{v_<} y) \wedge (\overline{\exists} s_{\{x,y\}}.\, [s_{\{x,y\}} \sim v_{R,\{1,2\}}]))\vee$$
$$((x >_{v_<} y) \wedge (\overline{\exists} s_{\{x,y\}}.\, [s_{\{x,y\}} \sim v_{R,\{2,1\}}])),$$

where we use $\overline{\exists} s_{\{x,y\}}.\, \varphi(s_{\{x,y\}})$ as an abbreviation standing for $\exists s_{\{x,y\}}.\, [s_{\{x,y\}} \in S \wedge N(s_{\{x,y\}}, B) = \{x, y\}]$.

Generally, we represent k-ary relation variables in a similar manner. That is, letting R^k be a k-ary relation, we think of R as a family of $k!$ collections of sets of k vertices from B. For each such k-set $\{x_1, \ldots, x_k\}$ there is a vertex $s_{\{x_1,\ldots,x_k\}}$ in S such that $N(s_{\{x_1,\ldots,x_k\}}, B) = \{x_1, \ldots, x_k\}$ (by G allowing a $(\varphi_0(x, \bar{y}_0), \varphi_1(x, \bar{y}_1), f)$-double powerset representation). Now consider a k-tuple $\bar{x} = (x_1, \ldots, x_k) \in R$, and let $\pi = \pi(\bar{x})$ be the permutation such that $x_{\pi(1)} \leq x_{\pi(2)} \leq \cdots \leq x_{\pi(k)}$ where in case of equality π maintains the original order (that is, if $x_i = x_j$ and $i < j$ then $\pi(i) < \pi(j)$). Let X_π be the collection of all k-tuples $\bar{x} \in R$ such that $\pi(\bar{x}) = \pi$. We represent X_π by a vertex $v_\pi \in V$ with the property that $N(v_\pi, S) = \{v_{\{x_1,\ldots,x_k\}} \mid (x_{\pi^{-1}(1)}, x_{\pi^{-1}(2)}, \ldots, x_{\pi^{-1}(k)}) \in R\}$ (such a vertex exists by our assumption on G). Finally, R is represented by the family $\{v_\pi\}$ where π ranges over all the permutations of $\{1, 2, \ldots, k\}$. This completes the proof. $\square$

Next we aim to show that a.a.s. $G(n, p)$ allows a $(\varphi_0(x, \bar{y}_0), \varphi_1(x, \bar{y}_1), f)$-double powerset representation with respect to two formulas in Q_{HAM} with

modified:2015-10-21
986 revision:2015-10-21_2

f growing to infinity quickly enough. The first formula defines a set of vertices all having the same degree. That is, $\varphi_0(x, v)$ will simply express the fact that $d(x) = d(v)$ (the details of writing φ_0 in $\mathcal{L}(Q_{\text{HAM}})$ are given in the beginning of the proof of Theorem 12 in Page 26). Thus, for an integer m we define $D(m) = \{v \in V \mid d(v) = m\}$. Somewhat abusing notation we also use $D(v)$ (for a given vertex v) for the set of all vertices with the same degree as v, that is, $D(v) = D(d(v))$. The second formula we shall use is even simpler. φ_1 will define a neighborhood of a vertex. That is, $\varphi_1(x, u) := x \sim u$. Lemma 18 shows that a.a.s. there are two vertices v, u such that $G(n, p)$ allows a $(\varphi_0(x, v), \varphi_1(x, u), f)$-double powerset representation with $f = \Theta(\log\log\log n)$.

The main idea of the proof of Lemma 18 is to show that $D(m)$ (with m properly defined) is pseudorandom, in a sense to be given soon. Then we shall use this pseudorandomness to show that there is a subset fitting to the requirement of Definition 16 (and every such subset is definable by $\varphi_1(x, u)$ for some vertex u).

We begin by picking a suitable degree. Let

$$m = np + h = np + \sqrt{n\left(\frac{1}{2}\ln n - \ln\ln\ln n + \alpha\right)2p(1-p)}, \qquad (2)$$

where α is $O(1)$.

Recall that $\varphi_0(x, v)$ is a formula in $\mathcal{L}(Q_{\text{HAM}})$ satisfying $G \models \varphi_0(x, v) \iff x \in D(v)$ and $\varphi_1(x, u) = x \sim u$.

Lemma 18. *Asymptotically almost surely there exist two vertices u, v in $G(n, p)$ such that $G(n, p)$ allows a $(\varphi_0(x, v), \varphi_1(x, u), f)$-double powerset representation where $f = \Theta(\log\log\log n)$.*

Before proving Lemma 18 we quote a fact regarding the distribution of $|D(m)|$ in $G(n, p)$. This distribution was studied a few times (e.g., [7, Chapter 3], [2] and [28]). For our application we cite (part of) Theorem 5.F of [2], which bounds the *total variation distance* between the distribution of the number of vertices of degree m and a Poisson distribution (The total variation distance between two probability measures λ and μ (both over the set of natural numbers) is defined as $\sup\{|\lambda(A) - \mu(A)||A \subset \omega\}$). We added the pace of decay of the total variation, originally appearing only in the proof, to the statement of the theorem.

modified:2015-10-21
986 revision:2015-10-21_2

Theorem 19 ([2, Theorem 5.F]). *Let W be a random variable counting the number of vertices of degree $m = m(n)$ in $G(n,p)$. Assume that np is bounded away from zero and that $(np)^{-1/2}|m - np| \to \infty$. Then the total variation distance between the distribution of W and a Poisson distribution with parameter $\lambda = n \cdot \Pr[\mathrm{Bin}(n-1,p) = m]$ tends to zero with n. Moreover, the total variation distance is bounded by*

$$\frac{Cm(m-np)}{np} \Pr[\mathrm{Bin}(n-1,p) = m],$$

where C is some constant independent of n.

Next we quote a Chernoff-type concentration theorem regarding the tails of a Poisson distribution. The last inequality in the statement is due to the fact that $\ln(1+x) \geq x - \frac{1}{2}x^2$ for $x > 0$.

Theorem 20 ([1, Theorem A.1.15]). *Let P be a random variable following a Poisson distribution with mean μ. Then for $\epsilon > 0$*

$$\begin{aligned} &\Pr[P \leq \mu(1-\epsilon)] \leq e^{-\epsilon^2\mu/2}, \\ &\Pr[P \geq \mu(1+\epsilon)] \leq \left[e^{\epsilon}(1+\epsilon)^{-(1+\epsilon)}\right]^{\mu} < e^{-\frac{1}{2}\epsilon^2\mu + \frac{1}{2}\epsilon^3\mu}. \end{aligned}$$

Proof of Lemma 18. By [7, Theorem 1.2] one has

$$\Pr[\mathrm{Bin}(n,p) = m] < C_1 \frac{\ln\ln n}{n}$$

For some constant C_1. On the other hand, by the counterpart [7, Theorem 1.5] we have

$$\Pr[\mathrm{Bin}(n,p) = m] > C_2 \frac{\ln\ln n}{n},$$

where C_2 is some positive constant. Hence we may write

$$\Pr[\mathrm{Bin}(n-1,p) = m] = \Theta(\ln\ln n/n). \tag{3}$$

We now argue that if we define the expectation $\mu = n \Pr[\mathrm{Bin}(n-1,p) = m]$ and set $\epsilon = \ln \mu/\sqrt{\mu}$, then for any constant K one has

$$\Pr[(1-\epsilon)\mu \leq |D(m)| \leq (1+\epsilon)\mu] \geq 1 - O((\log\log n)^{-K}). \tag{4}$$

modified:2015-10-21
986 revision:2015-10-21_2

Indeed, notice that m satisfies the conditions of Theorem 19 above. Therefore, the total variation distance between the distribution of $|D(m)|$ and a Poisson distribution with mean μ is bounded from above by $(Cm(m-np)/np)(\log\log n/n) = O((\sqrt{\ln n}\ln\ln n)/\sqrt{n})$.

By Theorem 20 and the fact that $\mu = \Theta(\ln\ln n))$ we have

$$\Pr[||D(m)|-\mu| \geq \epsilon\mu] \leq \Pr[|\mathrm{Pois}_\mu - \mu| \geq \epsilon\mu] + O\left(\frac{\sqrt{\ln n}\ln\ln n}{\sqrt{n}}\right) \leq$$
$$\leq \left(e^{-\epsilon^2\mu/2} + e^{-\epsilon^2\mu/2+\epsilon^3\mu/2}\right)(1+o(1)) \leq$$
$$\leq 3e^{-\frac{1}{2}\ln^2\mu} \leq 3\mu^{-\frac{1}{2}\ln\mu} = O\left(\left(\frac{1}{\log\log n}\right)^K\right)$$

for any constant K, as desired.

Let v be a vertex of degree m. By Lemma 22, a.a.s. $G(n,p)$ allows $\varphi(x,v)$-powerset representation. By Lemma 24, $G(n,p)[D(v)]$ is *pseudorandom* in the sense that the inner degrees and codegress behave as expected from a random graphs with edge density p, that is, the degrees are close (up to ϵ) to $p|D(m)|$ and the codegrees are close to $p^2|D(m)|$. Finally, the deterministic Lemma 25 tells us that every such pseudorandom graph has a fairly large subset B (logarithmic in the size of the graph) with the property that each of the subsets of B is the neighborhood in B of a vertex in the graph.

Let us recapitulate. We have seen that if v has degree m as above then

1. a.a.s. $|D(v)| = \Theta(\log\log n)$;
2. a.a.s. $G(n,p)$ allows $\varphi_1(x,v)$ powerset representation (Lemma 22) and
3. a.a.s. there is a vertex $u \in V[G(n,p)]$ such that $B = N(u, D(v))$ is of logarithmic size, and for every subset A of B there is a vertex $w \in D(v)$ such that $N(w, D(v)) = A$ (Lemma 24 and Lemma 25).

Therefore, a.a.s. $G(n,p)$ allows a $(\varphi_0(x,v), \varphi_1(x,u), f)$-double powerset representation with $f = \Theta(\log\log\log n)$. This completes the proof of the lemma. $\square$

The next fact is straightforward and easy: using second order logic (over sets) we can interpret arithmetic. While in terms of expressive power it is enough to have addition and multiplication, we define a few more relations for the benefit of readability.

modified:2015-10-21
986 revision:2015-10-21_2

Fact 21. *There exists a second order sentence*

$$\text{Arith} := \exists <^2, x_0, \ldots, x_{100}, +^3, \times^3, P, T, M.\, \varphi(<, x_0, \ldots, x_{100}, +, \times, P, T, M),$$

such that if a set B models Arith *then*

1. *$<$ is a linear order over the elements of B;*

2. *x_0 is the minimal element in the order, x_1 the second smallest element and so on up to x_{100};*

3. *$+, \times, P, T$ and M act as the standard addition, multiplication, base two power, base two tower[5] and modulo 100 relations.*

In addition, for every set B of cardinality at least 101, one has $B \models$ Arith.

Proving Fact 21 is fairly standard and not difficult. The proof may be found, e.g., in [30, Chapter 8].

Having arithmetic as above, that is, if $B \models$ Arith, we can express sentences dealing with the size of the set B. We shall use the following sentence saying that $0 \leq \log^* |B| < 50 \pmod{100}$.

$$\begin{aligned} \text{LogStar} := \exists x \in B.\, [&(\exists y \in B.\, T(x, y))\wedge \\ &(\forall z \in B.\, ((x < z) \rightarrow (\neg \exists y \in B.\, T(z, y))))\wedge \\ &M(x, 0) \vee M(x, 1) \cdots \vee M(x, 49)]. \end{aligned} \tag{5}$$

Now we have all the needed ingredients to prove Theorem 12.

proof of Theorem 12. Assume $p \leq 1/2$ (if $p > 1/2$ replace any $\sim$ by $\not\sim$). We begin by showing that given v, the set $D(v)$ is $\mathcal{L}(Q_{\text{HAM}})$-definable.

Let A and B be two $\mathcal{L}(Q_{\text{HAM}})$-definable sets (in our application both will usually be neighborhoods of vertices which are, of course, first order definable, and hence definable in any semiregular language). We say that $|A| \preceq |B|$ if there is a Hamilton cycle in the graph over $[n]$ having all the edges between $A \setminus B$ and $B \setminus A$ and all the edges with both endpoints in $V \setminus (A \setminus B)$. Clearly, $|A| \preceq |B|$ is expressible in $\mathcal{L}(Q_{\text{HAM}})$, provided that A and B are $\mathcal{L}(Q_{\text{HAM}})$-definable. Observe that per definition $|A| \preceq |B|$ if and only if $|A| \leq |B|$.

[5]The *base 2 tower* function $t(n)$ is defined by $t(0) = 1$ and $t(n) = 2^{t(n-1)}$ for all $n \geq 1$.

modified:2015-10-21
986 revision:2015-10-21_2

The last two paragraphs say that membership in $D(v)$, and hence also $\varphi_0(x, v)$, is expressible in $\mathcal{L}(Q_{\text{HAM}})$ since we may simply express $x \in D(v)$ or $\varphi_0(x, v)$ by writing $(|N(x)| \preceq |N(v)|) \wedge (|N(v)| \preceq |N(x)|)$.

At this point we have everything needed in order to apply Lemma 18 to encode any second order formula over a set of size $\log\log\log(n)$. We shall demonstrate the process for a specific formula having no limiting probability (also in the modular sense).

By Fact 21 there is a second order formula, Arith, that holds for every graph (actually, every set) with size at least 101 and lets us express arithmetic.

Consider the set $D(v)$ for some vertex v and a vertex $u \in V$. Using the encoding of Lemma 17 with the formulas $\varphi_0(v)$ and $\varphi_1(u)$, we encode the sentence Arith aforementioned into $\text{Arith}'(u, v)$. Similarly we encode the sentence LogStar from Equation (5) into a sentence $\text{LogStar}'(u, v)$. By Lemma 17 we have

$$N(u, D(m)) \models \text{Arith} \iff G(n, p) \models \text{Arith}'(u, v),$$

and similarly for LogStar and $\text{LogStar}'(u, v)$.

Let m be as in Equation 2 and let v be a vertex of degree m. By Lemma 18 there exists a vertex u such that $G(n, p)$ allows a φ_0, φ_1, f-double powerset representation with $f = \Theta(\log\log\log n)$. Therefore we have that a.a.s. $G(n, p) \models \text{Arith}'(u, v)$ for some vertices u and v. Now we may write the nonconverging sentence.

$$\begin{aligned}\text{NonConv} := \exists v, u.\, [&\text{Arith}'(u, v) \wedge \\ &(\forall v', u'.\, (\text{Arith}'(u, v) \rightarrow \neg(|N(u', D(v'))| \preceq |N(u, D(v))|))) \wedge \\ &\text{LogStar}'(u, v)].\end{aligned}$$

NonConv says that if we consider the maximal integer b for which there is a pair of vertices v, u such that $d(u, D(v)) = b$ and $\text{Arith}'(v, u)$ holds, then $0 \le \log^*(b) < 50 \pmod{100}$. By the above, we know that this maximal b satisfies $c \log\log\log n \le b$ for some constant c. We do not bother with the upper bound, rather we simply mention that trivially $b \le n$. This means that $\log^*(n) - 4 \le \log^*(b) \le \log^*(n)$.

Thus, if we consider an infinite sequence of numbers (n_i) all having $\log^*(n_i) = 49 \pmod{100}$, then $\lim_{i\to\infty} \Pr[G(n_i, p) \models \text{NonConv}] = 1$. On the other hand taking another sequence (n'_i) such that $\log^*(n'_i) = 99 \pmod{100}$

modified:2015-10-21
986 revision:2015-10-21_2

gives $\lim_{i\to\infty}\Pr[G(n_i',p)\models \text{NonConv}]=0$. Clearly the limiting probability of NonConv violates the modular convergence law of [23] as well.

A similar encoding may be applied for any second order formula. The proof is complete. $\square$

3.1 Three technical lemmas

Lemma 22. *Let $0<p<1$ be constant and consider $G(n,p)$. Then a.a.s for every set of vertices S of size $|S|\le C\log\log n$ (for some constant $C>0$) the following holds: For every subset $A\subset S$ there is a vertex $v\notin S$ such that*

$$N(v)\cap S=A.$$

Proof. Let $A\subset S$ and let $v\notin S$ be a vertex. The probability that $N(v)\cap S=A$ is exactly $p^{|A|}(1-p)^{\binom{|S|}{2}-|A|}\ge 2^{-c(C\log\log n)^2}$ where $c=\log(\min(p,1-p))$ is constant depending only on p. Therefore the probability that there is no witness for the set is bounded by $(1-2^{-c'(\log\log n)^2})^{n-C\log\log n}\le e^{-\sqrt{n}}$. Apply a union bound over all the $2^{C\log\log n}\le(\log n)^C$ possible subsets of S and then another union bound over all the $\sum_{k=1}^{C\log\log n}\binom{n}{k}\le\binom{n}{C\log\log n+1}\le e^{C'\log n\log\log n}$ sets of size $k\le C\log\log n$. Hence we have that a.a.s. for every set S of size at most $C\log\log n$ there is a witness for every subset, and the proof is complete. $\square$

Before stating and proving the nest lemma we cite yet another Chernoff type bound, this time regarding the tails of the hypergeometric distribution.

Theorem 23 ([21, Theorem 2.10])**.** *Let X be a random variable following a hypergeometric distribution with parameters n,l and m. Let $\mu=\mathbb{E}X=lm/n$ and $t\ge 0$. Then*

$$\Pr[X\ge\mathbb{E}X+t]\le e^{-t^2/(2(\mu+t/3))};$$
$$\Pr[X\le\mathbb{E}X-t]\le e^{-t^2/(2\mu)}.$$

Now we are ready to prove the following lemma, basically saying that degrees and codegrees inside $D(m)$ behave as expected.

Lemma 24. *Let $0<p<1$ be constant and let m be as above. Set $\epsilon=\ln\ln\ln n/\sqrt{\ln\ln n}$. Then a.a.s. for every $v\in D(m)$ one has*

$$d(v,D(m))=(1\pm\epsilon)p|D(m)|,$$

modified:2015-10-21
986 revision:2015-10-21_2

and a.a.s. for every pair of vertices u, v one has

$$|N(u) \cap N(v) \cap D(m)| = (1 \pm \epsilon)p^2|D(m)|.$$

Proof. Denote $S = \{v \mid d(v) = m\} = D(m)$. We are interested in typical degrees and co-degrees inside $G[S]$.

Choose $v \in [n]$. Denote $G_v = G[V \setminus \{v\}]$. Let $S_0 = \{u \in G_v \mid d_{G_v}(u) = m-1\}$ and $S_1 = \{u \in G_v \mid d_{G_v}(u) = m\}$. Then, by definition, $d(v, S) = d(v, S_0)$. Denote $X_0 = |S_0|, E[X_0] = \mu_0, X_1 = |S_1|, E[X_1] = \mu_1$. Notice that μ_0 and μ_1 are very close to each other since

$$\frac{\mu_0}{\mu_1} = \frac{(n-1)\binom{n-2}{m-1}p^{m-1}(1-p)^{n-2-(m-1)}}{(n-1)\binom{n-2}{m-2}p^{m-2}(1-p)^{n-2-(m-2)}} = \frac{(n-m)p}{(m-1)(1-p)} = 1 \pm O(1/\sqrt{n}).$$

Hereinafter equations containing X_i, μ_i or S_i are to be taken as two equations, one for $i = 0$ and one for $i = 1$.

By Equation (3) we have $\mu_i = \Theta(\log\log n)$. By Equation (4) we know that X_i is concentrated around its mean μ_i, that is, $\Pr[|X_i - \mu_i| > \epsilon\mu_i] < (\log\log n)^{-K}$ for any constant $K > 0$.

Now, $|S| = d(v, S_0) + (|S_1| - d(v, S_1))$ and, as already mentioned, $d(v, S) = d(v, S_0)$.

Consider the event A being "$v \in S \wedge |d(v, S) - p|S|| > \epsilon|S|$". First expose the degree of v, $\Pr[d(v) = m] = O(\log\log n/n)$. Condition now on $d(v) = m$, and expose G_v, with probability $1 - O((\log\log n)^{-K})$, the random variables X_0 and X_1 are close to their expectations and in particular are nearly equal.

Now, given $d(v)$ and the sets S_0, S_1, the random variables $d(v, S_0)$ and $d(v, S_1)$ are distributed hypergeometrically with parameters $n-1, X_i, m$. Let $\lambda_i = \mathbb{E}[d(v, S_i)] = mX_i/(n-1)$, and notice that $\lambda_i = (np+h)X_i/(n-1) = pX_i(1 \pm \epsilon) = p\mu_i(1 \pm 2\epsilon)$. Applying Theorem 23 we get

$$\begin{aligned}&\Pr[|d(v, S_i) - \lambda_i| \geq \epsilon\lambda_i] \leq e^{-(\epsilon\lambda_i)^2/(2\lambda_i(1+\epsilon/3))} + e^{-(\epsilon\lambda_i)^2/(2\lambda_i)} \leq \\ &\leq 3e^{-(\epsilon^2\lambda_i)/2} = O\left(\frac{1}{\log\log n}\right)^K,\end{aligned}$$

for every positive constant K.

All in all, the probability of the event A comes out to be $O(\log\log n/n) \cdot O((\log\log n)^{-K}) = o(1/n)$, and therefore a union bound over the vertex set is applicable and the first part of the lemma is proven.

986 revision:2015-10-21_2 modified:2015-10-21

We repeat the argument for codegrees. Fix two vertices u, v and consider $G_{u,v} = G \setminus \{u, v\}$. Let A be the event "$u, v \in S \wedge |\mathrm{codeg}(u, v, S) - p^2|S|| > \epsilon|S|$". We first expose the edge between u and v and assume w.l.o.g that u and v are not adjacent. Next we expose the degrees of u and v and with probability $\Theta((\log\log n/n)^2)$ we have that $d(u) = d(v) = m$ (these events being independent). Next we expose the edges of $G_{u,v}$ and look ar the sets S_0, S_1 and S_2 defined similarly to the above — $S_i = \{w \in G_{u,v} \mid d_{G_{u,v}}(w) = m - 2 + i\}$. This time we have

$$S = (N(u) \cap N(v) \cap S_0) \cup ((N(u) \triangle N(v)) \cap S_1) \cup (S_2 \setminus (N(u) \cup N(v))),$$

and thus

$$\begin{aligned} |S| =& \mathrm{codeg}(u, v, S_0) + (d(u, S_1) + d(v, S_1) - \mathrm{codeg}(u, v, S_1)) + \\ & (|S_2| - d(u, S_2) - d(v, S_2) + \mathrm{codeg}(u, v, S_2)). \end{aligned}$$

Again, we know that all the summands in the right hand side of the equation above are close to their expectation with high enough probability. That is, by Equation (4), with probability $(1 - O((\log\log n)^{-K}))$ the sets S_0, S_1 and S_2 are all of size $\Theta(\log\log n)$ and are all nearly of the same cardinality (the difference being of order $\epsilon \log\log n$). The degrees $d(u, S_i)$ and $d(v, S_i)$ follow a hypergeometric distribution with parameters $n - 2, |S_i|$ and m, and with probability $(1 - O((\log\log n)^{-K})$ they are all concentrated around their means. Finally notice that the codegree of u and v in S_i is determined by the degree of u inside $N(v) \cap S_i$, and thus it also follows a hypergeometric distribution with parameters $n - 2, p|S_i|(1 \pm \epsilon)$ and m, and the same hypergeometric tails bound applies. All in all we get that the probability of A is bounded by

$$O\left(\frac{\log\log n}{n^2}\right) \cdot O\left(\frac{1}{(\log\log n)^K}\right) = o(n^{-2}).$$

Applying the union bound over all pairs of vertices completes the proof. □

The next lemma deals with pseudorandom graphs, saying that if the degrees and codegrees of a graph are similar to those of a random graphs, then there is a small set of vertices such that every subset of it is induced by a vertex of the graph.

modified:2015-10-21
986 revision:2015-10-21_2

Lemma 25. *Let $0 < p < 1$ be constant. Let G be a graph on n vertices and let $\epsilon = n^{-b}$ where $0 < b < 1$ is some constant. Assume that G satisfies $\delta(G) \geq (p-\epsilon)n$ and $\Delta(G) \leq (p+\epsilon)n$. Additionally assume that for every two vertices u, v one has $(p^2-\epsilon)n \leq \operatorname{codeg}(u,v) \leq (p^2+\epsilon)n$. Then there is a constant $c = c(b,p) > 0$ and a set of vertices S of size $s = c\log n$ such that for every subset $A \subset S$ there is a vertex $v \in G$ having $N(v,S) = A$.*

Proof. Fix $0 \leq a \leq s$. Pick a set A of size a uniformly at random. Let X_A be the random variable counting the number of vertices v in G satisfying $A \subset N(v)$. We claim:
Claim 25.1: $\mathbb{E}[X_A] = np^a(1+\tilde{O}(n^{-b}))$ and $\operatorname{Var}[X_A] = \tilde{O}(np^a n^{1-b})$.
Proof of Claim 25.1: For every $v \in G$ let $X_{A,v}$ be the indicator random variable for the event $A \subset N(v)$. Let $d = d(v)$ be the degree of v. Then,

$$\Pr[A \subset N(v)] = \frac{\binom{d}{a}}{\binom{n}{a}} = \frac{d(d-1)\ldots(d-a+1)}{n(n-1)\ldots(n-a+1)}.$$

Since ($a = O(\log n)$ and)

$$\frac{d-a}{n-a} = \frac{d}{n}\left(1 + \frac{a}{n-a}\left(1-\frac{n}{d}\right)\right) = \frac{d}{n}\left(1+O\left(\frac{a}{n}\right)\right),$$

we have

$$\Pr[A \subset N(v)] = \left(\frac{d}{n}\right)^a\left(1+O\left(\frac{a^2}{n}\right)\right) = p^a(1\pm\epsilon)^a\left(1+O\left(\frac{a^2}{n}\right)\right).$$

Therefore,

$$\mathbb{E}X_A = \sum X_{A,v} = np^a(1\pm\tilde{O}(n^{-b})). \tag{6}$$

Next we wish to estimate $\operatorname{Var}[X]$. Consider two vertices u, v.

$$\mathbb{E}[X_{A,u}X_{A,v}] = \Pr[A \subset N(u) \wedge A \subset N(v)],$$

which is the probability of A being a subset of the common neighborhood of u and v. Let $d(u,v)$ be the codegree of u and v. Since $d(u,v)$ is bounded we get (similarly to the above):

$$\Pr[A \subset N^*(u,v)] = \frac{\binom{d(u,v)}{a}}{\binom{n}{a}} = \left(\frac{d(u,v)}{n}\right)^a\left(1+O\left(\frac{a^2}{n}\right)\right),$$

modified:2015-10-21
986 revision:2015-10-21_2

which yields

$$\mathbb{E}[X_{A,u}X_{A,v}] = p^{2a}(1 \pm O(\epsilon a)) = p^{2a}(1 \pm O(an^{-b})).$$

Now,

$$\begin{aligned}&\mathrm{Cov}[X_{A,u}, X_{A,v}] = \mathbb{E}[X_{A,u}X_{A,v}] - \mathbb{E}[X_{A,u}]\mathbb{E}[X_{A,u}] \leq \\ &p^{2a}(1 \pm O(an^{-b})) - (p^a(1 \pm O(an^{-b})))^2 = O(p^{2a}an^{-b}).\end{aligned}$$

Thus we have

$$\mathrm{Var}[X_A] \leq \mathbb{E}[X_a] + \sum_{u \neq v} \mathrm{Cov}[A, u, X_{A,v}] = np^a(1 + o(1)) + O(n^{2-b}p^{2a}a),$$

which gives the desired expression

$$\mathrm{Var}[X_A] = \tilde{O}(np^a n^{1-b})).$$

End of proof of Claim 25.1

Knowing $\mathbb{E}[X_A]$ and $\mathrm{Var}[X_A]$ we can apply Chebyshev's inequality. Let $\delta = n^{3c\log(1-p)}$ and recall that $0 < a < s = c\log n$. We may write

$$\begin{aligned}\Pr[|X_A - \mathbb{E}[X_A]| \geq & 3\delta\mathbb{E}[X_A]] \leq \frac{\mathrm{Var}[X_A]}{9\delta^2\mathbb{E}[X_A]^2} = O\left(\frac{n^{-6c\log(1-p)}np^a n^{1-b}}{n^2p^{2a}(1+o(1))}\right) = \\ =& O(n^{-6c\log(1-p)-b-c\log p}) = O(n^{-c'}),\end{aligned}$$

where $c' = b + c\log p + 6c\log(1-p)$.

Recall that by Equation (6) we have $\mathbb{E}[X_A] = np^a(1 \pm \tilde{O}(n^{-b}))$. If we require $c < b/(-3\log(1-p))$ we have $\delta = \tilde{\omega}(n^{-b})$ and thus

$$\begin{aligned}&\Pr[|X_A - np^a| \geq \delta np^a] \leq \\ &\Pr[|X_A - \mathbb{E}X_A| \geq (\delta + \tilde{O}(n^{-b}))np^a] \leq \\ &\Pr[|X_A - \mathbb{E}X_A| \geq 2\delta np^a] \geq \\ &\Pr[|X_A - \mathbb{E}X_A| \geq 3\delta\mathbb{E}[X_A]] = O(n^{-c'}).\end{aligned}$$

Call a set A having $|X_A - np^a| \geq \delta np^a$ "bad". Every set of size a is a subset of $\binom{n}{s-a}$ sets of size s. Hence the number of s-sets containing a bad

modified:2015-10-21
986 revision:2015-10-21_2

set is bounded by

$$\begin{aligned}
&\sum_{a=0}^{s}\binom{n}{a}O(n^{-c'})\binom{n}{s-a} = O(n^{-c'})\sum_{a=0}^{s}\binom{n}{a}\binom{n}{s-a} = \\
=&O(n^{-c'})\sum_{a=0}^{s}\binom{n}{s}\binom{s}{a} = O(n^{-c'})\binom{n}{s}\sum_{a=0}^{s}\binom{s}{a} = \\
&= O(n^{-c'})\binom{n}{s}2^s = O(n^{-c'+c})\binom{n}{s}.
\end{aligned}$$

Hence, requiring $c - c' < 0$ or $c < b/(1 - \log p - 6\log(1-p))$ gives that the number of s-sets having a bad subset is $o(\binom{n}{s})$.

Pick a set S of size s without any bad subset. Consider $A \subset S$, $|A| = a$. Let $W(A)$ be the set of vertices v such that $A \subset N(v)$ and let $W^*(A)$ be the set of vertices v such that $A = N(v)$. By the inclusion exclusion principle we have

$$\begin{aligned}
&|W^*(A)| = |W(A)| - \sum_{v\in S\setminus A}|W(A\cup v)| + \sum_{u\neq v\in S\setminus A}|W(A\cup u, v)| - \cdots = \\
=&np^a\sum_{k=0}^{s-a}(-1)^k\binom{s-a}{k}p^k(1\pm\delta) \geq \\
\geq&np^a\sum_{k=0}^{s-a}(-1)^k\binom{s-a}{k}p^k - \delta np^a\sum_{k=0}^{s-a}\binom{s-a}{k}p^k = \langle \text{ IE principle}^6 \rangle = \\
=&np^a(1-p)^{s-a} - \delta np^a\sum_{k=0}^{s-a}\binom{s-a}{k}p^k = \\
=&np^a(1-p)^{s-a} - \delta np^a\sum_{k=0}^{s-a}\frac{\binom{s-a}{k}p^k(1-p)^{s-a-k}}{(1-p)^{s-a-k}} \geq \\
=&np^a((1-p)^{s-a} - \delta/(1-p)^{s-a}) \geq np^a((1-p)^s - \delta/(1-p)^s).
\end{aligned}$$

[6] Assume you have $s-a$ biased coins that yield head with probability p. Let B be the event "No head when tossing all coins" (we assume independence of course). Define A_i as the event "The i'th coin gave head". Now, by the inclusion exclusion principle the probability of B is $\sum_{k=0}^{s-a}(-1)^k\binom{s-a}{k}p^k$. Direct computation gives that $\Pr[B] = (1-p)^{s-a}$.

modified:2015-10-21
986 revision:2015-10-21_2

Since $s = c\log n$ and $\delta = n^{3c\log(1-p)} = (1-p)^{3s}$ we have

$$\begin{aligned}|W^*(A)| \geq & np^a(1-p)^s(1+o(1)) \geq np^s(1-p)^s(1+o(1)) \geq \\ \geq & n^{1+c\log p + c\log(1-p)}.\end{aligned}$$

In particular, if we require $c < -1/(\log p + \log(1-p))$, we get that $W^*(A)$ is not empty and the proof is complete. ☐

Acknowledgment The first author thanks Michael Krivelevich for many helpful and instructive discussions regarding these problems in general, and the technical aspects of the proofs in particular.
Thanks also to Joel Spencer for introducing the problems and the beauty of this subject.

References

[1] N. Alon and J.H. Spencer. *The Probabilistic Method*. Wiley-Interscience Series in Discrete Mathematics and Optimization, Third edition, Aug. 2008.

[2] A.D. Barbour, L. Holst and S. Janson. *Poisson Approximation*. Vol 2 of *Oxford Studies in Probability*, Oxford University Press, Oxford, Mar. 1992.

[3] J.-M. L. Bars. Fragments of existential second-order logic without 0-1 laws. In *LICS*, pages 525–536, 1998.

[4] A. Blass, Y. Gurevich and D. Kozen. A zero-one law for logic with a fixed-point operator. *Information and Control*, 67(1-3):70–90, 1985.

[5] A. Blass and F. Harary. Properties of almost all graphs and complexes. *Journal of Graph Theory*, 3, 225–240, 1979.

[6] B. Bollobás. The chromatic number of random graphs. *Combinatorica*, 8(1):49–55, Mar. 1988.

[7] B. Bollobás. *Random graphs*. Number 73 in Cambridge Studies in avanced mathematics. Cambridge University Press, Cambridge, UK, Second edition, Jan. 2001.

modified:2015-10-21
986 revision:2015-10-21_2

[8] B. Bollobás. *Modern Graph Theory.* Graduate Texts in Mathematics. Springer, New York, Aug. 2002. Corrected Editon.

[9] K.J. Compton. 0-1 laws in logic and combinatorics. In R.I., ed., *Algorithms and Order*, volume 255 of *Advanced Study Institute Series C: Mathematical and Physical Sciences*, 353–383. Kluwer Academic Publishers, Dordrecht, 1989.

[10] R. Diestel. *Graph Theory.* No. 173 in Graduate Texts in Mathematics. Springer-Verlag, Heidelberg, 4th ed., Oct. 2010.

[11] H.D. Ebbinghaus and J. Flum. *Finite Model Theory.* Springer Monographs in Mathematics. Springer, Berlin, 2nd ed., 2006.

[12] H.-D. Ebbinghaus, J. Flum and W. Thomas. *Mathematical Logic.* Undergraduate Texts in Mathematics. Springer-Verlag, New York, 1984.

[13] P. Erdős and A. Rényi. On random graphs, I. *Publicationes Mathematicae*, 6:290–297, 1959.

[14] P. Erdős and A. Rényi. On the evolution of random graphs. In *Publication of the Mathematical Inst. of the Hungarian Academy of Sciences*, no. 5 in Acta Math. Acad. Sci. Hungar., pages 17–61, 1960.

[15] R. Fagin. Probabilities on finite models. *The Journal of Symbolic Logic* 41(1):50–58, Mar. 1976.

[16] Y.V. Glebskiĭ, D.I. Kogan, M.I. Liogon'kiĭ and V.A. Talanov. Range and degree of realizability of formulas in the restricted predicate calculus. *Cybernetics and Systems Analysis*, 5(2):142–154, Mar. 1969. (Russian original: Kibernetika 5(2):17–27, March-April 1969).

[17] S. Haber and M. Krivelevich. The logic of random regular graphs. *Journal of Combinatorics*, 1(3-4):389–440, 2010.

[18] S. Haber and S. Shelah. Failure of –1 law for sparse random graph, for strong logics, *Preprint.*

[19] S. Haber and S. Shelah. On a very expressive language obeying a zero-one law. *Preprint*, 2013+.

986 revision:2015-10-21_2 modified:2015-10-21

[20] S. Haber and T. Müller. On the logic of random geometric graphs. *Preprint*, 2013+.

[21] S. Janson, T. Łuczak and A. Ruciński. *Random Graphs.* Wiley-Interscience Series in Discrete Mathematics and Optimiziation. Wiley, New York, 2000.

[22] M. Kaufmann and S. Shelah. On random models of finite power and monadic logic. *Discrete Mathematics* 54(3):285–293, 1985.

[23] P.G. Kolaitis and S. Kopparty. Random graphs and the parity quantifier. In M. Mitzenmacher, ed., *STOC*, 705–714. ACM, 2009.

[24] P.G. Kolaitis and M.Y. Vardi. The decision problem for the probabilities of higher-order properties. In *Proceedings of the nineteenth annual ACM symposium on Theory of computing*, STOC '87, 425–435, New York, 1987. ACM.

[25] P.G. Kolaitis and M.Y. Vardi. 0-1 laws for fragments of existential second-order logic: A survey. In M. Nielsen and B. Rovan, editors, *MFCS*, vol. 1893 of *Lecture Notes in Computer Science*, 84–98. Springer, 2000.

[26] P. Lindström. First order predicate logic with generalized quantifiers. *Theoria*, 32(3):186–195, Dec. 1966.

[27] P. Lindström. On extensions of elementary logic. *Theoria*, 35(1):1–11, Apr. 1969.

[28] B.D. McKay and N.C. Wormald. The degree sequence of a random graph. I. The models. *Random Structures & Algorithms*, 11(2):97–117, Sept. 1997.

[29] S. Shelah and J. Spencer. Zero-one laws for sparse random graphs. *Journal of the American Mathematical Society*, 1(1):97–115, Jan. 1988.

[30] J.H. Spencer. *The Strange Logic of Random Graphs*. No. 22 in Algorithms and Combinatorics. Springer-Verlag, Berlin, 2001.

[31] V.A. Talanov. Asymptotic solvability of logical formulas. *Combinatorial-Algebraic Methods in Applied Mathematics*, 118–126, 1981. Gorky University, USSR. (Russian). Math Reviews 85i:03081.

modified:2015-10-21

986 revision:2015-10-21_2

[32] D.B. West. *Introduction to Graph Theory.* Prentice Hall, New Jersey, 2nd edition, Sept. 2000.

modified:2015-10-21

986 revision:2015-10-21_2