Spanning Structures and Undecidability in Random Graphs

A dissertation presented

by

Peter Cornelius Dolan

to

The Graduate School

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in

Mathematics

State University of New York

at Stony Brook

May 1989

State University of New York

at Stony Brook

The Graduate School

Peter Cornelius Dolan

We, the dissertation committee for the above candidate for the Doctor of Philosophy degree, hereby recommend acceptance of this dissertation.

Joel Spencer, Professor, Courant Institute of Mathematical Sciences, Dissertation Director

Bernard Maskit, Professor, Mathematics
Dissertation Co-director

Paul Kumpel, Professor, Mathematics
Chair of the Defense

Lowell Jones, Professor, Mathematics

Anita Wasilewska, Assistant Professor, Computer Science
Outside Member

This dissertation is accepted by The Graduate School

Alexander H. King

Abstract of the Dissertation

Spanning Structures and Undecidability in Random Graphs

by

Peter Cornelius Dolan

Doctor of Philosophy

in

Mathematics

State University of New York

at Stony Brook

1989

The main result of Chapter I is that if $n = \sum_{i=0}^{t} r^i$ and $p = c \log n/n$, where r is a chosen constant and c is a constant that depends on r, then there is a good algorithm that almost always finds an r-ary spanning tree in the random graph $G_{n,p}$. We also show that when np is constant or goes to infinity arbitrarily slowly, you can almost always find an r-ary tree on, respectively, n/k (k a constant) and n $(1-o(1))$ vertices. A result of de la Vega's on matchings that has been sharpened using martingales is used in this proof, which is also algorithmic. The existence of other large structures, particularly spanning trees of other shapes, is also investigated. A "99% solution" is given to a question of Erdős concerning the existence of a Hamiltonian cycle

with chords that join antipodal points.

In Chapter II we consider random graphs $G_{n,p}$ with $p = n^{-\alpha}$, $0 < \alpha < 1$, α rational. A graph property A holds almost always in $G_{n,p}$ if Pr [A holds in $G_{n,p}$] $\longrightarrow 1$ as $n \longrightarrow \infty$. We show that there is no decision procedure for determining if a general first order sentence S holds almost always in $G_{n,p}$.

This dissertation is dedicated to Margie and Neil, whose generosity made so much possible, and to my parents, who let me go my own way - even when it was the long way.

Table of Contents

Acknowledgements.....vii

Introduction..... 1

I. Large Structures in Random Graphs..... 3

II. Undecidable Statements and Random Graphs.....26

Bibliography.....43

ACKNOWLEDGEMENTS

First and foremost I would like to thank Joel Spencer for introducing me to combinatorics and helping me take my first halting steps as a mathematician. His enthusiasm for the subject and his dedication to teaching others about it will always serve as a model for me.

Most of the research in Chapter 1 was done, and a rebirth of my interest in mathematics occurred, while I was in Budapest on a Fulbright Fellowship. I would like to thank all the people who have worked to make the Fulbright program a success. The opportunity to live in the home of combinatorics may have meant the difference between finishing and not finishing this dissertation. Special thanks go to József Beck, Pál Erdős, and Vera T. Sós for their assistance and encouragement while I was in Budapest. Catherine Sevcenko of the United States Information Agency and László Fodor of the Hungarian Ministry of Education deserve mention for all of their help during that year. I would also like to thank the Mathematics Institute of the Hungarian Academy of Science and Eötvös Loránd University for their hospitality.

The Department of Discrete Mathematics at Adam Mickewicz

University in Poznań, Poland provided me with a pre-print of the paper by de la Vega that led to a major breakthrough in the spanning tree problem. I would also like to thank them for their hospitality.

For serving on my defense committee, and for their assistance in preparing my dissertation and my defense, I would like to thank Lowell Jones, Paul Kumpel, Bernard Maskit, and Anita Wasilewska

I am deeply indebted to Joan Conlin, Grace Dolan, Roberta Lauber, and Hugh Riley for their assistance in the typing and printing of this dissertation. They are in no way responsible for any errors contained in it.

Finally, I would like to thank the students, staff and faculty of the Department of Mathematics at Stony Brook who have been my friends, teachers, and co-workers over the past six years. I have tried to limit the personal acknowledgements to those people involved in the work on this dissertation. That work was done after four years of preparation and it was not done in isolation. To acknowledge properly all the people to whom I am indebted, either personally or professionally, would make this the longest chapter of my dissertation.

INTRODUCTION

In this dissertation we consider two problems in the theory of random graphs. The model we will use is the "probabilistic" one. $G_{n,p}$ will be a graph created randomly on a set of n vertices. The probability p that two vertices are joined by an edge will usually be a function of n. Each edge is chosen independently of the other edges. In the "static" model $G_{n,m}$, a graph is chosen from among all graphs on n vertices with m edges, with equal probability of choosing any such graph. Another model for a random graph is the "dynamic" model. Start with the empty graph on n vertices, and add randomly chosen edges one at a time until you have the complete graph on n vertices.

The relationships between the various models are discussed in [Bollobás 85] and [Palmer 85] . Asymptotically (as n tends to infinity), when $p = m/\binom{n}{2}$ the probabilistic and static models are close to identical. The probabilistic model is generally easier to work with from a computational point of view.

If A is a property of graphs then we say A holds *almost always* in $G_{n,p}$ if

$$\lim_{n\to\infty} \Pr[\ A \text{ holds for } G_{n,p}\ ] = 1 .$$

If the limit is zero then we say that A holds *almost never*. A function $p(n)$ is said to be a *threshold function* for A if

(i) $\lim_{n\to\infty} r(n)/p(n) = 0$ implies A holds almost never for $G_{n,r(n)}$

(ii) $\lim_{n\to\infty} r(n)/p(n) = \infty$ implies A holds almost always for $G_{n,r(n)}$

In the static model, if m is a function of n, then we can define threshold functions similarly. In the dynamic model we speak of the "hitting time" - the number of the edge whose addition almost certainly causes the graph to have property A. It should be noted that threshold functions are not unique. With the understanding that $q(n)$ is also a threshold function for A if $p(n) \sim q(n)$ (their ratio tends to a constant as n tends to infinity) , we will sometimes speak of "the" threshold function for A.

Some examples of threshold functions for various graph properties are:

Contains a k clique	$p(n) = n^{-2/(k-1)}$
Planarity	$p(n) = 1/n$
Connectivity	$p(n) = \log n/n$

Good introductions to the theory of random graphs can be found in [Spencer 87] and [Palmer 85] . [Bollobás 85] is a comprehensive reference.

Chapter I Large Structures in Random Graphs

The existence of subgraphs on a constant (that does not depend on n) number of vertices in $G_{n,p}$ has been exhaustively studied. The topic is given a complete treatment in Chapter 4 of [Bollobás 85]. On the other hand, little is known about thresholds for subgraphs whose order is a function of n. Arguments based on Chebychev's inequality (the "second moment method"), for example, can no longer be applied. When np is constant it has been shown that $G_{n,p}$ almost always contains long paths [Ajtai, Komlós, Szemerédi 81] , [de la Vega 79] , and any tree of constant maximum degree on $n/8$ vertices [de la Vega 83]. $np = \log n$ has been shown to be a threshold for the existence of a perfect matching [Erdős, Rényi 66] and a Hamiltonian cycle [Pósa 76]. In this chapter we investigate the existence of other "large" subgraphs in $G_{n,p}$, in particular those whose order is n or $(1-o(1))n$.

It should be noted that $p = \log n/n$ is the threshold for connectivity. Before this point $G_{n,p}$ almost always has more than one component and therefore no spanning subgraphs.

In this chapter ω is a function of n that tends to infinity arbitrarily slowly.

An r-ary tree is a branching process that starts from a single vertex. In each successive generation r new vertices are added to the existing terminal vertices. The main result of this chapter is that $p = \log n/n$ is a threshold function for the existence of an r-ary spanning tree for constant r.

THEOREM 1.1: Let $r \in \mathbb{N}$ be a constant greater than or equal to 2, $n = \sum_{i=0}^{t} r^i$, $p_0 = 8r/n$, $p_1 = \frac{c \log n + \omega}{n}$ where c is a constant that depends on r, and $\omega \longrightarrow \infty$ arbitrarily slowly. For

$$p = 1 - (1-p_0)(1-p_1) \sim p_0 + p_1 ,$$

the random graph $G_{n,p}$ almost always contains an r-ary spanning tree.

For a discussion on extending this theorem to $r = \log n$ and related questions, see the end of this chapter (pp. 20-25).

PROOF: We consider $G_{n,p}$ as the union $G_{n,p_0} \cup G_{n,p_1} = G_0 \cup G_1$. The edges of G_0 will be called the red edges and the edges from G_1 will be called the blue edges. (Note that a red edge and a blue edge joining the same pair of vertices are identified.) We will use an algorithm to find an r-ary tree of size $n/8$ whose edges are red. Then, by using matchings whose edges are blue, we will expand this core tree to a spanning tree. The proof requires two

lemmas:

Lemma 1.2: If $p = 8r/n$ then almost always $G_{n,p}$ contains an r-ary tree on $n/8$ vertices.

Sketch of Proof: This lemma is a consequence of a much more general result [de la Vega 83]. The proof uses a modified greedy breadth first search algorithm. The problem of searching for an r-ary tree in $G_{n,p}$ is converted into the problem of searching for an r-ary tree in the random family tree $F_{r+1,q}$ of a Galton-Watson process. (At time zero there is almost surely one initial point; each point of each generation spawns zero or $r+1$, not r , vertices with probabilities $1-q$ and q respectively. The extra "room" we get from the $r+1^{st}$ vertex is important.) ■

Lemma 1.3: If $p = \frac{\log n + \omega}{2n}$ then almost always the bipartite random graph $G_{n,n;p}$ contains a perfect matching.

Proof: We follow [Bollobás 85]. The original result is in [Erdős and Rényi 66].

Let V_1 and V_2 be the vertex classes of $G_{n,n;p}$ and for $A \subset V_i$ $(i = 1,2)$ let $\Gamma(A)$ be the set of all vertices adjacent to some vertex in A.

At this edge probability $G_{n,n;p}$ almost always has no isolated vertices. Suppose there is no complete matching.

Claim: There must be a set $A \subset V_i$ $(i = 1,2)$ such that:

i) $|\Gamma(A)| = |A| - 1$

ii) $A \cup \Gamma(A)$ is connected

iii) $2 \leq |A| \leq (n+1)/2$

Proof of Claim: If there is no complete matching, then by Hall's theorem there is some set $A \subset V_i$ such that $|\Gamma(A)| < |A|$. If we choose such a set A of smallest cardinality, then conditions (i) and (ii) hold since otherwise A could be replaced by a proper subset of itself. Without loss of generality, suppose $A \subset V_1$. $B = V_2 \setminus \Gamma(A)$ is such that $\Gamma(B) \subset V_1 \setminus A$, and so

$$|\Gamma(B)| \leq |V_1| - |A| < |V_2| - |\Gamma(A)| = |B|$$

and $|A| \leq |B| = |V_2| - |\Gamma(A)| = n - (|A| - 1)$, so $|A| \leq (n+1)/2$. ∎

Let F_a be the event that there is a set A of order a that satisfies the above three conditions and $n' = \lfloor (n+1)/2 \rfloor$. We show that $\sum_{a=2}^{n'} \Pr[F_a] = o(1)$.

Let $A_1 \subset V_1$, $A_2 \subset V_2$ and $|A_1| = |A_2| + 1 = a$, then the probability that the graph spanned by $A_1 \cup A_2$ has at least $2a - 2$ edges (is connected) and no vertex of A_1 is joined to a vertex in $V_2 \setminus A_2$ is at most

$$\binom{a(a-1)}{2a-2} p^{2a-2} (1-p)^{a(n-a+1)}$$

There are $2\binom{n}{a}$ choices for A with $|A| = a$, and $\binom{n}{a-1}$ choices for $\Gamma(A)$, therefore

$$\sum_{a=2}^{n'} \Pr[F_a] \le 2 \sum_{a=2}^{n'} \binom{n}{a} \binom{n}{a-1} \binom{a(a-1)}{2a-2} \left(\frac{\log n + \omega}{2n}\right)^{2a-2}$$

$$(1-p)^{a(n-a+1)}$$

$$\le 2 \sum_{a=2}^{n'} \left(\frac{en}{a}\right)^{a} \left(\frac{en}{a-1}\right)^{a-1} \left(\frac{ea}{2}\right)^{2a-2} \left(\frac{\log n}{n}\right)^{2a-2} n^{-a + a^2/2n}$$

$$\le \sum_{a=2}^{n'} (e \log n)^{3a} n^{1-a+a^2/n} = o(1) \qquad \blacksquare$$

Proof of Theorem 1.1: We will take $r \ge 8$. A slight modification (embedding technique, see Theorem 1.8), is used for smaller values. $\sum_{i=0}^{t-1} r^i \le n/8$, so by Lemma 1.2, we almost always find a red r-ary tree on $\sum_{i=0}^{t-1} r^i$ vertices. Let C be the set of vertices of this red tree. Let E be the set of terminal vertices in C. Partition the vertices of $G \setminus C$ into r sets $B_1, \ldots, B_r$; $|B_i| = r^{t-1}$. We then look at the blue edges. It is important to realize that the blue edges are independent of the red edges. By Lemma 1.3, since the number of B_i's is constant, the blue edges almost always give us perfect matchings from E to each of the B_i's. The

edges of these matchings are the remaining edges of the spanning tree.

Finally, we examine the value of the constant c. $n = \alpha r^t$ where $1 \le \alpha \le 2$. Thus $|B_i| = n/\alpha r$, so in LEMMA 1.3 $c = \alpha r/2$. When $r < 8$, $\sum_{j=0}^{t-1} r^j > n/8$. In this case, the embedding technique used in THEOREM 1.8 should be used. This will reduce the value of the constant in this case since the matchings used in the first growth steps will be between sets of size $n/2$ rather than $n/\alpha r^2$. ■

Two things should be noted here. The blue edges will almost always give us perfect matchings among the B_i's too. This can be used to show the existence of other large structures in $G_{n,p}$. It is also important to note that de la Vega's method is essentially a greedy algorithm and that good algorithms exist to find matchings [Lovász, Plummer 86]. This proof provides a good method for finding specific large structures in a random graph.

At $p = c/n$, c a constant, or ω/n , where $\omega \to \infty$ and is $o(\log n)$, the random graph $G_{n,p}$ is not connected but it can be shown to contain large trees of size n/c' and $n(1-o(1))$ respectively. (It should also be noted that if $x \in \mathbb{R}$, then at $p = \{\log n + x + o(1)\}/n$, the probability that $G_{n,n;p}$ contains

a perfect matching tends to $e^{-2e^{-x}}$ [Bollobás 85].) We define a *defect r-ary tree* to be a tree formed by deleting terminal vertices from an r-ary tree if $r \geq 8$. When $r < 8$ we also allow deletion of vertices (along with all their descendents) from any of the last three generations.

THEOREM 1.4: If $np = \omega$, then $G_{n,p}$ almost always contains a defect r-ary tree on $(1-o(1))\ n$ vertices.

PROOF: The proof is identical to the proof of THEOREM 1.1 except that the edges of G_1 will not give us perfect matchings. By sharpening a result of de la Vega's [de la Vega 82], we show that the probability that a matching will fail to cover $(1-o(1))\ n$ vertices is exponentially small (THEOREM 1.5). This will be sufficient since the number of matchings used is constant. ■

THEOREM 1.5: If $np = \omega$, then the probability that the greedy algorithm defined in the proof fails to find $n - 2n/\omega$ independent edges in the random bipartite graph $G_{n,n;p}$ is less than $2e^{-n/2\omega^2}$.

PROOF: The proof follows [de la Vega 79] with the with the addition of a martingale argument to show that the probability that more than $o(n)$ vertices are not covered by the matching produced

by the algorithm is exponentially small. A martingale is a sequence of random variables $X_0, \cdots, X_n$ for which $E[X_{i+1}|X_i] = X_i$. An example of a martingale is a coin flip game in which the player starts with X_0 dollars and flips a coin n successive times. If the coin comes up heads he wins a dollar. If it is tails he loses a dollar. X_i is the amount of money he has after the ith flip. We give two proofs of this theorem, both of which use the following result on the deviation of X_n:

Lemma: 1.6: Let $X_0, \cdots, X_n$ be a martingale with $|X_{i+1} - X_i| \leq c$. Then

$$\Pr[\,|X_n - X_0| > \lambda c \sqrt{n}\,] < 2\exp\{-\lambda^2/2\}$$

Proof: [Shamir, Spencer 87] . ■

Proof I of **Theorem** 1.5: Let $G_{n,n;p} = A \cup B$ be a bipartite random graph. $A = \{a_1, \ldots, a_n\}$ and $B = \{b_1, \ldots, b_n\}$. Think of the edges of the graph as being hidden [Matula 87]. We expose the vertices of B and the edges that join them to A one vertex at a time. At the start of the algorithm all vertices of A are available. If b_i is joined to at least one available vertex of A, then from them choose the one of least index, say a_k . a_k is no longer available and we add $\{b_i, a_k\}$ to the matching. If b_k is not joined to any available vertices then it is not covered by the matching.

In this case too we delete the vertex of least index from the set of available vertices. Actually, as we expose the vertices of B and the edges that join them to A, the remainder of the graph can still be treated as a random graph. So in either case it does not matter how we select the vertex a_k which will no longer be available.

After j vertices have been exposed the probability that b_{j+1} is not matched with a vertex in A is $(1 - p)^{n-j}$. Thus the expected number of vertices of B not covered by the matching is

$$\sum_{j=0}^{n-1} (1 - p)^{n-j} < \sum_{k=1}^{\infty} (1 - p)^{k} = 1/p - 1$$

which is $o(n)$.

Let Z be the total number of vertices not covered by the matching, and $Z_i = E[Z \mid \text{exposure of the first } i \text{ vertices of B}]$; $Z_0 = E[Z]$ and $Z_n = Z$. The sequence $Z_0, \cdots, Z_n$ forms a martingale. The central idea is that after we have uncovered k vertices, the *average* of the expected number of vertices not covered after looking at the edges from the k+1st vertex is Z_k . Suppose we have uncovered k vertices of B and $u(k)$ of them have not been matched. Let $F(k)$ be the expected number of unmatched vertices in one vertex class when this algorithm is applied to $G_{k,k;p}$ (so $F(n) = E[Z] = Z_0$).

$$\begin{aligned} E[Z_{i+1}|Z_i] &= \{E[u(i+1)] + F(n-i-1)\} \mid Z_i \\ &= u(i) + (1-p)^{n-i} + F(n-i-1) \\ &= u(i) + F(n-i) \\ &= Z_i \end{aligned}$$

Furthermore, since $u(i+1) = u(i)$ or $u(i)+1$, $|Z_{i+1} - Z_i| \leq 1$ and the bound of LEMMA 1.6 can be applied to the deviation of Z with $c = 1$. Setting $\lambda = \sqrt{n}/\omega$, we have that $\Pr[Z > 2n/\omega] < e^{-n/2\omega^2}$. (This method of forming a martingale can be used in various situations. General conditions for when a sequence of conditional expectations will form a martingale can be found in [Feller 66]. For applications to the chromatic number of a random graph, see [Bollobás 88] and [Shamir, Spencer 87].) ∎

PROOF II of THEOREM 1.5: In the algorithm used in the first proof we always delete a vertex from the set of available vertices, regardless of whether or not vertex b_i is matched. This was done to make calculations easier. If we delete only those vertices that are matched from the set of available vertices, the sequence $Z_0,\ldots,Z_n$ will still form a martingale. It becomes more difficult to show that $|Z_{i+1} - Z_i| \leq 1$.

We will use the following notation:

M(G) = the number of matchings when the algorithm is applied to G

H_i = the total exposed portion of G after exposing the edges incident with b_i

K_{i+1} = the restriction of the random graph to $\{b_{i+1}\} \cup A$

K^* = a fixed graph on $\{b_{i+1}\} \cup A$

After exposing b_{i+1}, the exposed portion of G is $H_i \cup K_{i+1}$. Define $\phi(G)$ to be identical to G with the exception that the edges incident to b_i are changed to match K^* .

$$Z_i = \sum_{G \supset H_i} M(G) \Pr[G \mid H_i]$$

$$= \sum_{K_{i+1}} \sum_{\substack{G \supset \\ H_i \cup K_{i+1}}} M(G) \Pr[G \mid H_i \wedge K_{i+1}] \Pr[K_{i+1}]$$

$$= \sum_{K_{i+1}} \Pr[K_{i+1}] \sum_{\substack{G \supset \\ H_i \cup K_{i+1}}} M(G) \Pr[G \mid H_i \wedge K_{i+1}]$$

$$Z_{i+1} = \sum_{\substack{G \supset \\ H_i \cup K^*}} M(G) \Pr[G \mid H_i \wedge K^*]$$

Because the unexposed portion of the graph can be treated as a random graph, $\Pr[G \mid H_i \wedge K_{i+1}] = \Pr[\phi(G) \mid H_i \wedge K^*]$ so

$$Z_i - Z_{i+1} = \sum_{K_{i+1}} \Pr[K_{i+1}] \sum_{\substack{G \supset \\ H_i \cup K_{i+1}}} [M(G) - M(\phi(G))] \Pr[G \mid H_i \wedge K_{i+1}]$$

To show $|Z_i - Z_{i+1}| \leq 1$, we need to show

$$|M(G) - M(\phi(G))| \leq 1 .$$

Suppose we have a bipartite graph on sets $\{c_1, \ldots c_k\}$ and

$\{d_1, \ldots d_l\}$. Apply the matching algorithm. Now change the vertex of least index adjacent to d_l, either by deleting or adding an edge to the graph. This will change the output of the matching algorithm. d_l will be matched with a new vertex. If, in the original graph, that vertex had been matched with vertex d_i , then d_i will now be matched with its "second choice". That vertex, if any, will not be available to vertices of index greater than i . If d_i was matched in the original graph, and if the vertex it was paired with would have been the first choice of another vertex d_j, then d_j will now be matched with its first choice. If d_j was matched in the original graph, then the vertex it was paired with becomes available to vertices of index greater than j . It is possible that chain reactions will occur. One of the chains will either add one to the number of pairs matched, or leave it unchanged. The other chain will either decrease the number by one, or leave it unchanged. Therefore the net change will be either ±1 or 0 .

While this is a better algorithm than the one used in the first proof, the best bound that can be given for the number of vertices in B that do not get paired is the same as the bound for the first algorithm. ■

Corollary 1.7: If $np = c$ a constant, then $G_{n,p}$ almost always con-

tains a defect r-ary tree on n/c' vertices, c' a constant.

Proof: Follow the proof of **Theorem** 1.4, replacing ω with the constant value of np. ■

The existence of trees whose growth from the core is sufficiently uniform can be shown at $p = \frac{c \log n + \omega}{n}$:

Theorem 1.8: Suppose $k_1, k_2 \in \mathbb{R}$ and $m \in \mathbb{N}$ are constants greater than 1, and $\{T^1, T^2, \ldots, T^n, \ldots\}$ is a sequence of trees, each on n vertices. If each tree can be expressed as a sequence of distinct trees: $T_0^n \subset T_1^n \subset \cdots \subset T_j^n = T^n$, $j \leq m$ such that:

i) $|T_0^n| \leq n/8$

ii) for all $x \in T^n$, $\delta_0(x) \leq r$ and $\delta_{i+1}(x) \leq \delta_i(x) + 1$, all i ($\delta_i(x)$ is the degree of x in T_i^n),

iii) $n/k_1 \leq |T_j^n \setminus T_{j-1}^n| \leq n/k_2$

then, for $p = \frac{c \log n + \omega}{n}$, c a constant, $T^n \subset G_{n,p}$ almost always.

Proof: The proof is uses the idea of **Theorem** 1.1 . Replace G_1 with a sequence of j random graphs G_{n,p_i} , $p_i = \frac{c_i \log n + \omega}{n}$, c_i constant. The edges of these graphs are used to build the desired tree step by step.

If $|T_{i-1}| < n/2$, then partition the vertices of G into two disjoint sets A and B of $n/2$ vertices with $T_{i-1} \subset A$. Otherwise set $B = G \setminus T_i$ and let A be a set of $|B|$ vertices in T_{i-1} that includes all the vertices of T_{i-1} whose degree increases as we go from T_{i-1} to T_i . The edges of G_{n,p_i} almost always contain a matching from A to B. Edges of the matching are used to enlarge T_{i-1} to T_i . The third condition in the statement of the theorem ensures that the threshold for these matchings is $\log n/n$. ■

The starfish with r arms (a central vertex with r paths of length n/r radiating from it) does not satisfy the above criteria. However it can be shown that the threshold for the appearance of this graph is also $p = c \log n/n$.

THEOREM 1.9: If $n \equiv 1 (\mathrm{mod}\, r)$, then when

$$p = (r/n)\{\log n + \log\log n + \omega\}$$

$G_{n,p}$ almost always contains a starfish with r arms whose center is any vertex of the graph.

PROOF: Partition the vertices into r sets that are disjoint except for one point that is common to all of them. Almost always, each of these sets will contain a Hamiltonian cycle [Pósa 76]. From each cycle delete one of the vertices adjacent to the common point. ■

It should be noted that if r is an even number the constant can be improved. Partition the vertices into $r/2$ sets and delete the middle edge from each Hamiltonian cycle. Other "starfish type" structures can be created by deleting edges other than the middle edge or an edge adjacent to the center.

Among trees of constant maximum degree r , the r-ary tree and the r-starfish have the minimum and the maximum diameters respectively. These techniques seem to break down for certain graphs with diameter in the middle of the range of possible diameters. For example, take an r-ary tree of depth k and to each terminal vertex adjoin a path of length r^k . The threshold for a matching between two sets of size $n^{1/2}$ or the existence of a Hamiltonian cycle in a set of that size is $p \sim \log n/n^{1/2}$, which is well beyond the threshold for connectivity.

It seems likely that the actual threshold for any spanning tree of constant maximum degree is $\log n/n$. More precisely, if $\{T^1, T^2, \ldots, T^n, \ldots\}$ is a sequence of trees, each on n vertices, with constant maximum degree, then it is likely that for $p = c \log n/n$, $T^n \subset G_{n,p}$ almost always. There are two reasons to believe this. The first is that the thresholds for the appearance of the $\log n$-ary spanning tree and the rumor tree (see page 20) are both at most $(\log n)^2/n$ (THEOREM 1.11, COROLLARY 1.13). Both

of these trees have maximum degree $\log n$. The second is that in the above example we could take S to be the set of all terminal vertices of the r-ary part together with the remaining vertices. The order of S is roughly $n - n^{1/2}$, so the threshold for the existence of a Hamiltonian cycle in this set is $\log n/n$. The terminal vertices would be randomly placed around that cycle; the distances between them should be close to r^k. By deleting edges as in the starfish example, we will have a tree that is close to what we want.

One of the difficulties in proving that $\log n/n$ is the correct threshold is that as the tree you are interested in requires more conditions to describe it, you are trying to take tighter and tighter control over a random process. The following example, which involves an attempt to meld two structures that both have threshold $\log n/n$ - a Hamiltonian cycle and a one factor - further illustrates this.

Erdős asks for the threshold function for the appearance of a Hamiltonian cycle with $n/2$ chords that join antipodal points of the cycle. We give a "99% solution".

Theorem 1.10: If $p_0 = \dfrac{\log n + \log\log n + \omega}{n}$ and $p_1 = \dfrac{50 \log n + \omega}{n}$ and $p = 1 - (1 - p_0)(1 - p_1) \sim (c \log n + \log\log n + \omega)/n$ then

$G_{n,p}$ almost always contains a Hamiltonian cycle with $n/2$ chords that join each vertex to a point that is at most $n/100$ vertices away from its antipodal point on the cycle.

Proof: Define G_0 and G_1 as in theorem 1.1. G_0 almost always contains a Hamiltonian cycle [Pósa 76]. Partition the vertices of the cycle into segments of length $n/100$. By lemma 1.3, the edges of G_1 almost always give complete matchings between all pairs of antipodal segments. (Note that the number of segments is constant. If an arbitrary constant c had been used instead of 100, then the 50 in p_1 would be replaced with $c/2$.) ∎

Consider a matching between two segments A and B given by the edges of G_1 . If there is more than one matching, then choose one at random. Orient the cycle and number the vertices of the two segments from 1 to $n/100$. If we define $\sigma(x)$ for $x \in A$ to be the point in B that x is matched with then $\begin{bmatrix} 1 & \cdots & n/100 \\ \sigma(1) & & \sigma(n/100) \end{bmatrix}$ is a random permutation of the set $\{1,\ldots,n/100\}$. The number of fixed points tends to a Poisson distribution with mean 1. A fixed point of this permutation corresponds to an antipodal chord. Thus we expect 50 antipodal chords among the edges of those matchings.

In actuality the edges of G_1 should create more antipodal chords. Since the probability that a given pair of vertices is

joined by an edge is $p_1 = (50 \log n + \omega)/n$, the number of antipodal chords formed by edges of G_1 has a binomial distribution $B(n/2, p_1)$ This has mean $(50 \log n + \omega)/2$. However, we may not specify that these are edges of the matchings in the proof of Theorem 1.10; this is a *random* graph. If we were to delete the edges of antipodal chords in G_1, there might no longer be complete matchings between the pairs of segments.

By looking at matchings among different subsets of the vertices of the cycle and changing the value of the constant 50 it can be shown that Hamiltonian cycles with different chordal structures exist in $G_{n,p}$. In particular, the edges of G_1 almost always give matchings between *all* pairs of segments in THEOREM 1.10. While it may not be possible to show the existence of a specific chord pattern it may be possible to show the existence of an approximation as was done here. It is also possible to show that a given graph exists almost always when the edge probability is of the order of magnitude $\log n/n$ by showing that the graph can be embedded in a graph that is known to exist almost always, and then deleting edges as in THEOREM 1.9.

So far we have only considered trees of constant maximum degree. Define T_k, a *rumor tree* of depth k, inductively:
T_0 is a single vertex.

T_{k+1} is formed from T_k by having each vertex of T_k spawn a single new vertex, so $|T_k| = 2^k$.

Attempts to determine a threshold function for the existence of a spanning rumor tree in $G_{n,p}$ led to the results in this chapter. Unfortunately there is still a gap between the point below which it is known that there is almost always no such tree (the threshold for connectivity) , and the next theorem.

THEOREM 1.11: If $n = 2^k$ and $p = \frac{c(\log n)^2}{n}$, then $G_{n,p}$ almost always contains a spanning rumor tree.

PROOF: We follow the proof of THEOREM 1.8. Set $p_0 = n^{-1}$ so that G_0 will contain a rumor tree of depth k, where $k \in \mathbb{N}$ is a constant [Bollobás 85] . We successively double the size of the tree in G_0 using edges from $G_1, \ldots, G_{t-k}$, $t \sim \log n$, with $p_j \equiv \frac{\log n + \omega}{n}$ as in the proof of THEOREM 1.1 . Let T_j be the tree formed from the edges in $G_0 \cup \cdots \cup G_j$. Partition the vertices of G into $A \cup B$, $|A| = |B| = n/2$ with $T_j \subset A$. The edges of G_{j+1} contain a matching from A to B. Restrict this matching to T_j to form T_{j+1}.■

This technique of embedding the existing tree in one set of a bipartition of G is useful in sharpening the constant in THEOREM 1.1 when $r < 8$, and in the construction of other trees.

The maximum degree of the rumor tree is $\log_2 n$. de la Vega's theorem in its original form cannot be used to generate a core rumor tree of size n/c. His use of the law of large numbers relies on the maximum degree being constant. To extend LEMMA 1.2 to trees whose maximum degree is a function of n requires a second moment calculation. Furthermore, he bounds a random variable F with a convergent infinite series that diverges when you square the summands. The crucial observation is that if you do get to depth t in the search algorithm in the $(r+1)$-ary family tree, you stop because that tree will contain the tree that you are searching for. While asymptotically this does not improve THEOREM 1.11 (the net result is that the square term is multiplied by $(1 - \frac{1}{\log\log n})$), this may be useful in the case of other spanning structures.

The following discussion refers to de la Vega's paper on trees in sparse random graph [de la Vega 83].

Suppose you are searching for a $(\log n)$-ary tree. The depth t of this tree is less than $\log n/\log\log n$. Thus,

$$E(F^2) < \frac{1}{1-q_\infty}\left[1 - q + \sum_{k=1}^{t} a_k\,(r+1)^{2k+2}\right]$$

$$< 1 + (r+1)^2 \sum_{k=1}^{t}\left[\frac{r+1}{2}\right]^k .$$

Since $\sum_{k=1}^{t} \left[\frac{r+1}{2} \right]^k \le 2 \left[\frac{r+1}{2} \right]^t$ we can bound the variance of F:

$$\mathrm{Var}(F) \le E(F^2) \le 1 + (r+1)^2\, 2 \left[\frac{r+1}{2} \right]^t .$$

This in turn bounds the variance of U_k:

$$\mathrm{Var}(U_k) \le \frac{n_{r,k}}{r} \left[1 + (r+1)^2\, 2 \left[\frac{r+1}{2} \right]^t \right].$$

$$E(U_k) \le n_{r,k} \left(1 + \frac{2r^{-2}}{3} \right)$$

$$n_{r,k} \left(1 + 2r^{-2}\right) = E(U_k) \left[1 + \frac{1}{3r^2+2} \right]$$

$$\Pr\left[U_k \ge E(U_k) \left[1 + \frac{1}{3r^2+2} \right]\right]$$

$$\le \frac{(3r^2+2)^2\, n_{r,k} \left[1 + (r+1)^2\, 2 \left[\frac{r+1}{2} \right]^t \right]}{r\, n_{r,k}^2 \left[1 + \frac{2}{3r^2} \right]}$$

$$\sim \frac{c\, r^5\, 2 \left[\frac{r}{2} \right]^t}{n}$$

$$= \frac{c \log^5 n \; \frac{\log n}{\log\log n} \; (\log n)^{\log n / \log\log n}}{n\; 2^{\log n / \log\log n}}$$

$$= \frac{c \log^6 n}{n^{1/\log\log n}} \longrightarrow 0 \text{ as } n \longrightarrow \infty .$$

Following de la Vega, the number of vertices U_k' visited

before a tree of depth k is found is bounded above by the sum of U_k and three other terms:

i)the number of free vertices that might have been visited

ii)the number, say W, of vertices visited on eventual first dying trees

iii)unused "reserved" vertices of θ_{k-1}

The first term is $5n$, and the last term is bounded by $n_{r,k-1}$. W is the sum of a random number X of independent random variables distributed as F. X has a geometric distribution with parameter q_∞ . Therefore the expected number of "bad starts" is $E[X] = (1-q_\infty)/q_\infty = c_1$, and $E(W)$ is $E(X)E(F) \leq c_1 r$.

Since r is a function of n we need a second moment calculation. Recall that while W is the product of two random variables this does not mean $Var(W) = Var(X)\ Var(F)$. The relationship is more complex [Feller 66] :

$$Var(W) = E(X)Var(F) + Var(X)E^2(F)$$

Recall that X has a geometric distribution with parameter q_∞ so $Var[X] = (1-q_\infty)/q_\infty^2 = c_2$. As above, $Var[F]$ can be bounded by $E[F^2]$. A similar second moment calculation as for U_k works for U_k' since the terms that do not involve F^2 are not asymptotically significant.

This gives us the following extension of LEMMA 1.2:

COROLLARY 1.12: If $p = c\lfloor \log n \rfloor / n$ then $G_{n,p}$ almost always contains a $\lfloor \log n \rfloor$-ary tree on $n/8$ vertices. ■

These second moment calculations required that the depth of the tree be at most $\log n / \log\log n$ and this requirement must be included in any generalization of de la Vega's result, which is for all trees of a given maximum degree. This depth restriction is the reason why there is no asymptotic improvement in THEOREM 1.10 .

In terms of THEOREM 1.1 the sets B_i will have order roughly $n/\log n$. This will force us to set

$$p_1 \approx \log n (\log n - \log\log n)/n$$

so we have

COROLLARY 1.13: If $p = c\lfloor \log n \rfloor^2 / n$, then $G_{n,p}$ almost always contains a $\lfloor \log n \rfloor$-ary spanning tree. ■

The language of The First Order Theory of Graphs consists of a countable number of variables, Boolean connectives, existential and universal quantifiers, equality, and adjacency (denoted $I(x,y)$). The axioms are $\forall x\ \neg I(x,x)$ (no loops) and $I(x,y) \Rightarrow I(y,x)$. (Note that implication can be expressed in terms of "and" and "or".) Some examples of things we can say are:

There is a path of length 3

$$\exists x,y,z,w\ I(x,y) \wedge I(y,z) \wedge I(z,w) .$$

There are no isolated points

$$\forall x\ \exists y\ I(x,y) .$$

Every triangle is contained in a K_4

$$\forall x,y,z\ \{[I(x,y) \wedge I(y,z) \wedge I(x,z)] \Rightarrow \exists w\ [I(x,w) \wedge I(y,w) \wedge I(z,w)]\} .$$

Many graph properties cannot be expressed in this language. For example you cannot say that the graph is connected, planar, or Hamiltonian. When looking at random graphs it is particularly important to keep in mind that you cannot express the order of the graph in a first order statement.

At first glance it might seem that problems in the first order theory of graphs would not be difficult. This is not true.

For example, it is not known whether there exists a finite graph G on two or more vertices that satisfies the following conditions:

i) G is triangle free.

$\neg\ [\exists\ x,y,z\ \ I(x,y) \wedge I(x,z) \wedge I(y,z)]$

ii) For any pair of vertices x,y there exists a vertex z that is adjacent to x and not adjacent to y.

$\forall x,y\ \exists z\ \ I(z,x) \wedge \neg I(z,y)$

iii) For any three vertices x,y,z , if x is not adjacent to y, then there exists a vertex w that is adjacent to x and y but is not adjacent to z

$\forall x,y,z\ \ \neg I(x,y) \Rightarrow [\exists w\ \ I(x,w) \wedge I(y,w) \wedge \neg I(z,w)]$

In this chapter we will consider random graphs $G_{n,p}$, $p = n^{-\alpha}$, α a rational number between 0 and 1 . The main result of this chapter is that there is no decision procedure that separates those first order statements that hold almost always in $G_{n,p}$ from those whose negation holds almost always. This is slightly stronger than showing that there is no decision procedure that determines whether or not a first order statement holds almost always.

To prove the main theorem we begin by looking a the neighborhoods N_i of fixed sets of vertices in $G_{n,p}$. Subsets $\tilde{N}_i \subset N_i$ are defined inductively. A vertex is in $\tilde{N}_i$ if it is adjacent to

a vertex in $\tilde{N}_{i-1}$. We then give a procedure for defining a graph on the vertices of the last $\tilde{N}_i$. It is shown that for any finite graph H there is almost always a choice of fixed sets of vertices so that the graph we define is isomorphic to H (*Universal Representation Theorem, 2.4*). For any first order statement A about a finite graph, a corresponding first order statement A^* is shown to hold almost always in $G_{n,p}$ if A holds for some finite graph, and never if A holds for no finite graph. If a decision procedure as described in the previous paragraph existed for $G_{n,p}$, then the correspondence between A and A^* would allow us to use that procedure to separate those A that hold for some finite graph from those that hold for no finite graph. This would contradict the *Trachtenbrot - Vaught Theorem* (THEOREM *2.5*) [Trachtenbrot 50], which says that there is no decision procedure that separates those first order statements that hold for some finite graph from those that hold for no finite graph.

We will first consider the case of $\alpha \leq 1/2$. When $\alpha > 1/2$ we need to make some modifications in the proof. These will be explained at the end of the chapter. Fix disjoint sets X_i of r_i vertices (r_i is defined below). The exclusive neighborhood of X_i, $N_i = N_i(X_i) \subset G$, is defined by:

$$y \in N_i \text{ if } \begin{cases} I(y,x) & \text{for all } x \in X_i \\ \neg I(y,z) & \text{for all } z \in X_j \ , j \neq i \end{cases}$$

We define $\tilde{N}_i \subseteq N_i$ by:

(1) $\tilde{N}_1 = N_1$

(2) $z \in \tilde{N}_i$, $1 < i$, if there exists $w \in \tilde{N}_{i-1}$ such that $I(w,z)$ and $z \in N_i$

Note $\tilde{N}_i \subset N_i$ for all i

We define a first order predicate MEM : A vertex y satisfies $\mathrm{MEM}(y;\ X_1,\ldots,X_a)$ if and only if $y \in \tilde{N}_a$. It is important to realize that X_i is shorthand for $x_{i1}, \ldots, x_{ir_i}$. This is necessary if MEM is to be a first order predicate. We cannot talk about "the subset X_i" in the first order language.

Theorem 2.1: There is a sequence of integers $r_1,\ldots,r_a$ such that for fixed subsets $X_1,\ldots,X_a$ of $G_{n,p}$, $|X_i| = r_i$, the order of the set of vertices that satisfy the first order predicate $\mathrm{MEM}(y;\ X_1,\ldots,X_a)$ tends to a Poisson distribution with mean 1.

Proof: We will assume $\alpha = a/b$ is in lowest terms.

Put:
$$r_1 = \left\lfloor \frac{b}{a} \right\rfloor$$
$$d_1 = b(\mathrm{mod}\ a)$$
$$r_i = \begin{cases} r_1 & \text{,if } d_{i-1} + k > a \\ r_1 - 1 & \text{,else} \end{cases}$$
$$s_i = b - r_i a$$
$$d_i = d_{i-1} + s_i - a$$

We will show that the orders of the sets defined above are

$$|N_i| \sim n^{s_i/b} \text{ , and } |\tilde{N}_i| \sim n^{d_i/b} \text{ .}$$

We first examine the distribution of $|N_i|$. The probability that a vertex is in N_i is $p^{r_i} \prod_{i \neq j} (1-p^{r_j}) \sim p^{r_i}$ and this is independent over the vertices of G. Therefore,

$$|N_i| \longrightarrow B(\ n\ ,\ p^{r_i}\ )$$

$$\Rightarrow\ E[\ |N_i|\ ] = \mu_i \longrightarrow n^{s_i/b} \text{ .}$$

The probability that a binomial random variable is too far from its expectation is exponentially small [Chernoff 52]:

$$\Pr[\ |\ |N_i| - \mu_i\ | > \lambda\sigma\ ] < 2\ e^{-2\lambda^2}$$

where $\sigma^2 = np(1-p)$ is the variance of the binomial random variable (Note that for the edge probabilities we are considering $(1-p) \sim 1$) . This gives us:

$$\Pr[\ |\ |N_i| - \mu_i\ | > \mu_i^{3/4}\ ] < 2\ e^{-2\mu_i^{1/2}}$$

Since there is constant number a of N_i's , the probability that the order of some of them differ from their expectations by more than $\mu_i^{3/4}$ is less than $2\sum e^{-\mu_i^{1/2}} < 2a\ e^{-\min\{\mu_i^{3/4}\}} \longrightarrow 0$.

Example: If $p = n^{-5/23}$ then we expect $\sim n^{3/23}$ vertices in all of the N_i except for N_3 where we expect $\sim n^{8/23}$.

$$\Pr[\ |\ |N_i| - \mu_i\ |\ > \mu_i^{3/4}\ ] < \begin{cases} 2e^{-2n^{4/23}} & , i = 3 \\ 2e^{-2n^{3/46}} & , \text{else} \end{cases}$$

The probability that the order of some of the N_i's differ from their expectations by more than $\mu_i^{3/4}$ is less than $10\ e^{-n^{3/46}}$. o

This bound on the tail of the binomial distribution is actually stronger than we need. In this case, because we are taking the sum of a constant number of independent random variables, the bound

$$\Pr\ [\ |\ X - \mu\ |\ > \lambda\sigma\ ] < 1/\lambda^2$$

given by Chebychev's inequality [Rényi 70] would have been sufficient.

We now examine $|\tilde{N}_i|$. $\tilde{N}_1$ is defined to be N_1, whose order was bounded in the above calculations. If $z \in N_{i+1}$ then, by induction, the number of points in $\tilde{N}_i$ adjacent to z is

$$B(\ (1-o(1))n^{d_i/b}\ ,\ n^{-a/b}\ )$$

So we expect $(1-o(1))n^{(d_i-a)/b}$ adjacent vertices in $\tilde{N}_i$. It helps to view the edges of G as "covered". We first exposed the edges from the X_i to the rest of the graph. Now we are successively exposing the edges from N_i to N_{i+1}. This allows us to treat the restriction of G to the edges between N_i to N_{i+1} as a random bigraph [Matula 87]. Thus, the probability that z has no

neighbors in $\tilde{N}_i$ is

$$(1 - n^{-a/b})^{(1-o(1))n^{d_i/b}}$$

$$\sim 1 - (1-o(1))n^{(d_i-a)/b}$$

so the probability that z has neighbors in $\tilde{N}_i$ is

$$\sim 1 - (1 - (1-o(1))n^{(d_i-a)/b})$$

$$= (1-o(1))n^{(d_i-a)/b}$$

Call a point in N_{i+1} good if it is joined to at least one point in $\tilde{N}_i$. Take $\tilde{N}_{i+1}$ to be the set of all good points in N_{i+1}. Since the edges joining two distinct points of N_{i+1} to $\tilde{N}_i$ are all independent the number of good points in N_{i+1} is

$$B(\ (1-o(1))n^{s_{i+1}/b},\ (1-o(1))n^{(d_i-a)/b}\ )$$

Thus, we expect $(1-o(1))n^{d_{i+1}/b}$ vertices in $\tilde{N}_i$, and, by the same arguments as for $|N_i|$, the probability that the actual number of good vertices differs from the expected number by more than a factor of $1+o(1)$ is exponentially small.

Example: When $p = n^{-5/23}$, $|N_1| \sim |N_2| \sim n^{3/23}$; $|\tilde{N}_1| \sim n^{3/23}$
If $x \in N_2$ then the number of vertices in $\tilde{N}_1$ that are adjacent to x tends to

$$B(\ (1-o(1)n^{3/23}\ ,\ n^{-5/23}\ )$$

$$\Pr[x \text{ has no neighbors in } \tilde{N}_1] \sim (1 - n^{-5/23})^{(1-o(1))n^{3/23}}$$

$$\sim 1 - (1-o(1))n^{-2/23}$$

$$\Pr[\ x \in \tilde{N}_2\ ] \sim (1-o(1))n^{-2/23}$$

$$|\tilde{N}_2| \sim B(\ (1-o(1))n^{3/23}\ ,\ (1-o(1))n^{-2/23}\ )$$

Following the arguments used for $|N_i|$, the probability is exponentially small that the difference between the actual and the expected number $n^{1/23}$ of vertices in $\tilde{N}_2$ is more than $o(n^{1/23})$. o

We claim that $|X_a| = r_1$ and $|\tilde{N}_{a-1}| \sim (1+o(1))n^{(a-k)/b}$, which will imply $|\tilde{N}_a| \sim 1$ (d_a will equal zero). This is a consequence of the following algebra lemma:

Lemma 2.2: If $a,b \in \mathbf{Z}^+$ with $b > a$, $(b,a) = 1$, and d_i defined as follows:

$d_1 = b(\text{mod } a)$, $d_i = (d_{i-1} + d_1)(\text{mod } a)$, $i > 1$,

then $d_a = 0$.

Proof: $d_i = d_{i-1} + b - r_i a - a = (d_{i-1} + b)(\text{mod } a)$

$= (d_{i-1} + d_1)(\text{mod } a)$. $(a,b) = 1$ implies $(a,b(\text{mod } a)) = 1$.

Therefore, d_1 generates the cyclic group of order a. ■

Example: See figure 1, page 34. o

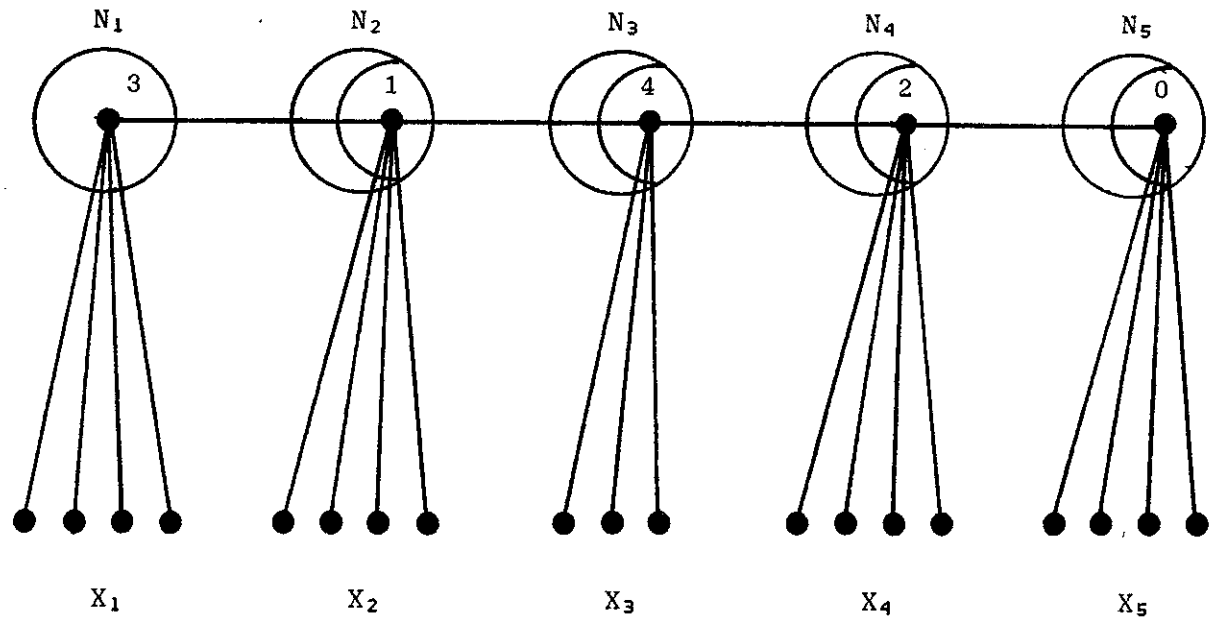


Figure 1

$$p = n^{-5/23}$$

The number in the circle is d_i,

where $|\tilde{N}_i| \sim n^{d_i/23}$

If we have a sequence X_n of binomial random variables $B(n,p(n))$ with $np \longrightarrow \lambda$ a constant, then the sequence of binomial distributions converges to a Poisson distribution with mean λ. Following [Rényi 70] :

$$\Pr[X_n = k] = \binom{n}{k} p^k (1-p)^k = \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right)$$

$$\sim \frac{\lambda^k}{k!} \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

$$\sim \frac{\lambda^k e^{-k}}{k!}$$

Since, by LEMMA 2.2, the expectation for the order of $\tilde{N}_a$ tends to 1, $|\tilde{N}_a| \longrightarrow P(1)$. ■

Let $H(X_1,..,X_t)$ be the set of vertices y that satisfy $MEM(y\ ;\ X_1,...,X_t)$

COROLLARY 2.3: For any natural number k, there almost always exist disjoint sets of vertices $X_1,...,X_t$ in $G_{n,p}$ with the property that $|H(X_1,...,X_t)| = k$.

PROOF: Fix k and $q > 0$. Choose disjoint sets of vertices $X_1,...,X_{a-1}$, X_a^1 ,..., X_a^q with $|X_i| = r_i$ and $|X_a^j| \equiv r_1$. Define N_i , N_a^j as in the main theorem. The sets N_a^j will be disjoint from each other and from the other N_i . The same argu-

ments as in the proof of THEOREM 2.1 apply to the orders of these sets. Put $H_i = H(X_1,\ldots,X_a^i)$. For each i the probability that $|H_i| = k$ tends to $1/ek!$. Therefore, the expected number of H_i that contain k vertices tends to $q/ek!$. Because we have independence (the N_a^j are disjoint) the number of H_i that contain exactly k vertices has a binomial distribution. Therefore the probability that none of the H_i contain exactly k vertices tends to $(1- 1/ek!)^q$. Since q is arbitrary, this probability can be made arbitrarily small. ∎

We actually have more. Since the neighborhoods are independent of each other, the probability that we have fewer than $q/2ek!$ H_i's of order k is less than $exp\{-q/2ek!\}$

We now define a first order predicate $ADJ(u,v;\ X_1,\ldots,X_a,A)$, abbreviated $ADJ(u,v)$, on the pairs of vertices in H . This will induce a graph on the vertices of H . As with MEM , the definition of ADJ must be modified for $\alpha > 1/2$. These modifications will be explained at the end of the chapter.

When $\alpha \leq 1/2$, fix a set A of $r_a - 2$ additional vertices (A may be empty). For the pairs of vertices $\{u,v\}$ in H we look at the disjoint neighborhoods of the sets $\{u,v\} \cup A$ as we did for the N_i in THEOREM 2.1 . The number of vertices in each of these

will have the same distribution as N_a . $ADJ(u,v)$ holds if and only if there are vertices in the neighborhood of $\{u,v\} \cup A$ that are adjacent to vertices of $\tilde{N}_{a-1}$.

The number of edges from N_a to the neighborhood of $\{u,v\} \cup A$ tends to a poisson distribution with mean 1. For two vertices u,v $Pr[\sim ADJ(u,v)]$ tends to $1/e$. Furthermore, these probabilities are mutually independent among all pairs $\{u,v\}$ because we used disjoint sets to define them. We have now defined a random graph with edge probability tending to $(1-1/e)$ on the vertices of H by $I(u,v)$ if and only if $ADJ(u,v)$.

THEOREM 2.4 (***Universal Representation Theorem***) : For any finite graph H on v vertices there almost always exists a collection of disjoint sets $X_1,\ldots, X_a$, A in $G_{n,p}$ such that the random graph defined on $H(X_1,\ldots,X_a)$ by $ADJ(u,v;\ X_1,\ldots,X_a,A)$ is isomorphic to H .

PROOF: Fix $X_1,\ldots,X_{a-1}$, A . Let the sets X_a^j be q disjoint sets of r_1 vertices and define H_j as in the proof of COROLLARY 2.3 . If the order of H is v, and E is the event that the graph defined on H_j by ADJ is isomorphic to H, then, by the comment preceding the theorem,

$$\Pr[\ E \mid |H| = v\ ] \longrightarrow \frac{n!}{|\mathrm{Aut}(H)|} \left(1 - \frac{1}{e}\right)^{l} \left(\frac{1}{e}\right)^{\binom{v}{2}-l} = \hat{p},$$

which is a positive *constant*.

Since the H_j are disjoint, the number of graphs isomorphic to H has a binomial distribution $B(q\ ,\ \frac{1}{ev!}\ \hat{p}\ )$. The probability that there is no graph isomorphic to H is $\left(1 - \frac{1}{ev!}\ \hat{p}\right)^{q}$, which can be made arbitrarily small by taking q large. ■

Example: If H is a cycle on four vertices then

$$\hat{p} = 3\ \left(1 - \frac{1}{e}\right)^{4} \left(\frac{1}{e}\right)^{2} .$$ o

To prove our main result we need the following theorem:

THEOREM 2.5: (*Trachtenbrot - Vaught Theorem*) There is no decision procedure that separates those first order statements A that hold for some finite graph from those A that hold for no finite graph.

Proof: [Trachtenbrot 50] . ■

Theorem 2.6: There is no decision procedure that separates those first order statements A that hold almost always for the random graph $G_{n,p}$ from those for which $\neg A$ holds almost always.

PROOF: We have shown that for all finite graphs H

$$\lim_{n\to\infty} \Pr[\ \exists\ X_1,\ldots,X_a,A \subset G_{n,p} : H(X_1,\ldots,X_a;A) \cong H\ ] \longrightarrow 1$$

For any first order statement A about finite graphs define A^* to be the following statement about the random graph $G_{n,p}$:

$$\exists\ X_1,\ldots,X_a,A \subset G_{n,p} \text{ such that } A \text{ holds for } H(X_1,\ldots,X_t;A)$$

EXAMPLE: Take the statement "there are no isolated vertices in the graph" to be A. In the first order language this is expressible as: $\forall\ y\ \exists\ z\ I(y,z)$. The corresponding statement A^* for the random graph with $p = n^{-2/5}$ would be:

$\exists\ (x_1,x_2,x_3,x_4)\ [\forall y\ \mathrm{MEM}(y;\ x_1,x_2,x_3,x_4) \Rightarrow \exists\ z\ \mathrm{MEM}(z;\ x_1,x_2,x_3,x_4)$
$\wedge\ \mathrm{ADJ}(y,z)]$

where we have defined the two first order predicates MEM and ADJ as follows

$$\mathrm{MEM}(y;x_1,x_2,x_3,x_4) \Leftrightarrow \exists\ z\ I(z,x_1) \wedge I(z,x_2) \wedge I(y,x_3) \wedge I(y,x_4) \wedge I(z,y)$$

$$\mathrm{ADJ}(u,v) \Leftrightarrow \mathrm{MEM}(u;\ x_1,x_2,x_3,x_4) \wedge \mathrm{MEM}(v;\ x_1,x_2,x_3,x_4) \wedge \exists\ t\ \mathrm{MEM}(t;\ x_1,x_2,u,v)$$

o

Since

$$\lim_{n\to\infty} \Pr[\ A^* \text{ holds in } G_{n,p}\ ] = \begin{cases} 0 \text{ if } A \text{ holds for no finite graph} \\ 1 \text{ if } A \text{ holds for some finite graph} \end{cases}$$

a decision procedure that could separate those statements that hold almost always in $G_{n,p}$ from those whose negation holds almost always would allow us to separate those A that hold for some finite graph from those that hold for no finite graph. This would contradict the *Trachtenbrot - Vaught Theorem.* ∎

EXAMPLE: Earlier, we stated a problem in the first order theory of graphs (p.27). If such a decision procedure existed, then we could use it to determine whether or not a finite graph (not a random graph) exists that satisfies conditions (i), (ii), and (iii). It is important to realize that Theorem 2.6 does not imply that this problem is undecidable. While that may be the case, all we know is that the question has not been decided.

When $\alpha > 1/2$ we need to redefine N_i slightly because in some cases X_i will be empty. View the process dynamically. For those i where X_i is empty we take N_i to be all of G except for those vertices in the N_j's corresponding to nonempty X_j's or in some $\tilde{N}_k$, $k < i$. The number of vertices of G that we omit is small compared to the size of G so it does not affect the final calculations except for a finite number of additional $o(1)$ terms.

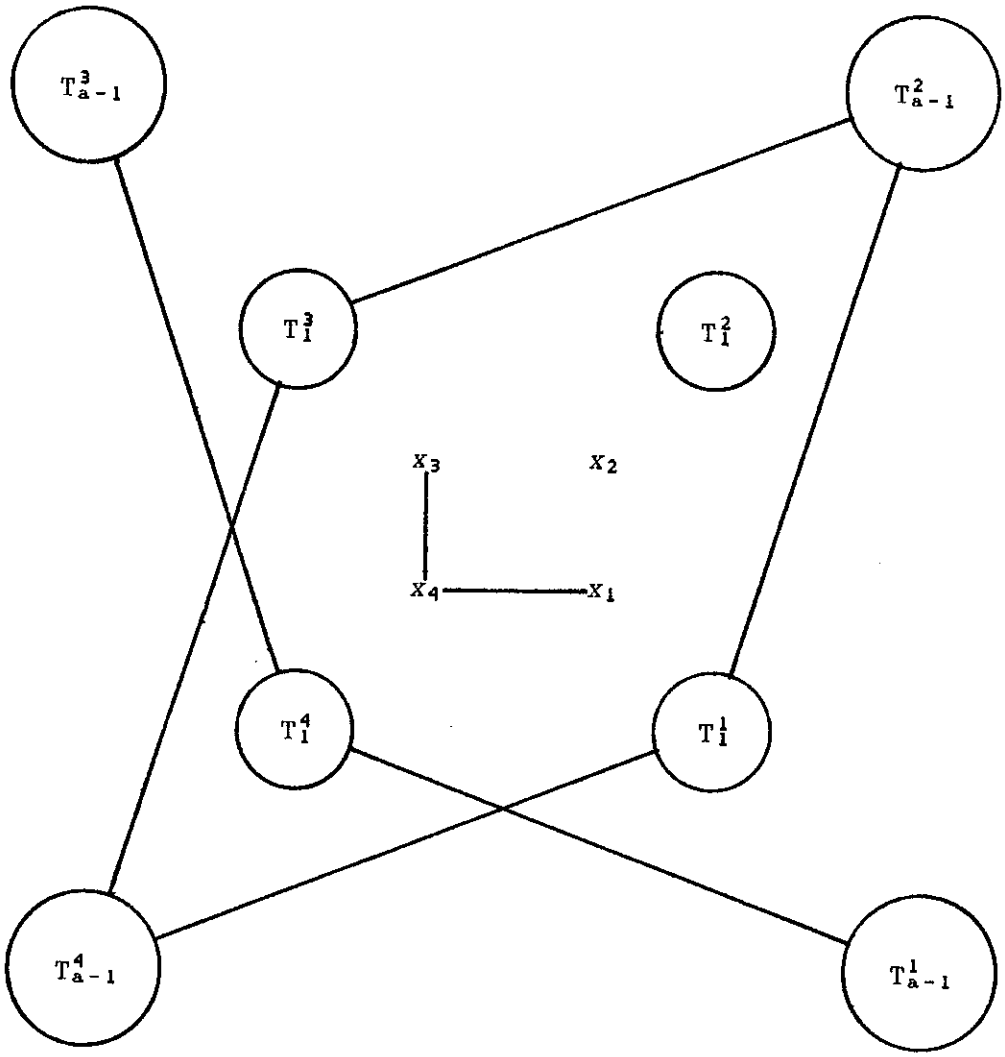
EXAMPLE: If $p = n^{-5/7}$ then $X_2 = X_4 = \emptyset$ so we put $N_2 = G \setminus N_1$,

$N_4 = G \setminus (N_1 \cup \tilde{N}_2 \cup N_3)$. We expect $n^{2/7}$ vertices in N_1 , N_3, N_5 .

$$\Pr[\ |\ |N_i| - \mu_i\ | > n^{6/28}] < 2e^{-n^{1/7}}\ ,\ i = 1,\ 3,\ 5$$

The probability that the order of any of these N_i's differs from its expectation by more than a factor of $1+o(1)$ is less than $6e^{-n^{1/7}}$. o

For $\alpha > 1/2$ (and therefore $|X_a| = 1$) we also must redefine ADJ . We will work in N' , the set of vertices that are not in any of the sets X_i , N_j for those $X_j \neq \emptyset$, or $\tilde{N}_k$. $|N'| \sim |N|$. For each vertex x_i in an H_k of order v , fix a family of sets $S_2^i,\ldots,S_{a-1}^i$, with $|S_j^i| = |X_j|$. Letting $S_1^i = \{x_i\}$ we form sequences of sets $\tilde{T}_1^i,\ldots,\tilde{T}_{a-1}^i$ in the same way that we formed $\tilde{N}_1,\ldots,\tilde{N}_{a-1}$, using the S_j^i's in the place of the X_i's. We say $x_i \longrightarrow x_j$ for $x_i, x_j \in H_k$ if and only if there are edges from $\tilde{T}_1^i$ to $\tilde{T}_{a-1}^j$. ADJ(u,v) holds if and only if $u \longrightarrow v$ and $u \longleftarrow v$. (See figure 2, page 43.) We need the relationship ADJ to be symmetric. This is why we require both $u \longrightarrow v$ and $u \longleftarrow v$. Since all the $\tilde{T}$ sets are disjoint and the number of edges between a $\tilde{T}_1$ and a $\tilde{T}_{a-1}$ has a Poisson distribution with mean tending to 1 (by the same calculation as for N_a), the only change in the calculations of THEOREM 2.4 is that the random graph defined by ADJ has edge probability $(1-1/e)^2$.



$H_k = \{x_1, x_2, x_3, x_4\}$

A line joining two sets indicates the existence of edges joining some of their vertices. If there is no line joining two sets, then there are no edges from one to the other.

In this situation, the indicated graph will be induced on H_k .

Figure 2

BILBIOGRAPHY

Ajtai, M., Komlós, J., Szemerédi, E., The longest path in a random graph, *Combinatorica 1*, 1-12, 1981

Bollobás, B., *Random Graphs*, Academic Press, New York, 1985

Bollobas, B., The chromatic number of random graphs, *Combinatorica 8*, 49-56, 1988

Chernoff, H., A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Annals Math. Stat. 23*, 493-507,1952

Erdős, P., Rényi, A., On the existence of a factor of degree 1 of a connected random graph, *Acta. Math. Acad. Sci. Hungar. 17*, 359-368

Feller, W., *An Introduction to Probability Theory and Its Applications Vol. II*, Wiley, New York, 1966

Grimmett, G.R., Stirzaker, D.R., *Probability and Random Processes*, Oxford University Press, New York, 1982

Lovász, L., Plummer, M., *Matching Theory*, Akadémiai Kiadó, Budapest, 1986

Matula, D., Expose-and-merge exploration and the chromatic number of a random graph, *Combinatorica* 7, 275-284, 1987

Palmer, E.M., *Graphical Evolution*, Wiley, New York, 1985

Pósa, L., Hamiltonian circuits in random graphs, *Discrete Math. 14*, 359-364, 1976

Rényi, A., *Probability Theory*, Akadémiai Kiadó, Budapest, 1970

Shamir, E., Spencer, J., Sharp concentration of the chromatic number on random graphs $G_{n,p}$, *Combinatorica* 7, 121-129, 1987

Spencer, J., *Ten Lectures on the Probabilistic Method*, Society for Industrial and Applied Mathematics, Philadelphia, 1987

Trachtenbrot, B.A., Impossibility of an Algorithm for the Decision Problem on Finite Classes, *Doklady Akademiia Nauk. S.S.R. 70*, 1950

de la Vega, W.F., Long paths in random graphs, *Studia Sci. Math. Hungar. 14*, 335-340, 1979

de la Vega,W.F., Sur la cardinalité maximum des couplages d'hypergraphs aleatoires uniformes, *Discrete Math. 40*, 315-318, 1982

de la Vega, W.F., Trees in sparse random graphs, *Journal of Combinatorial Theory, Series B 39*, 77-85, 1988