ON SPARSE GRAPHS WITH DENSE LONG PATHS

by

P. Erdös
R. L. Graham
E. Szemerédi

STAN-CS-75-504
SEPTEMBER 1975

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

ON SPARSE GRAPHS WITH DENSE LONG PATHS

by

P. Erdös*, R. L. Graham**, E. Szemerédi*
Stanford University
Stanford, California

INTRODUCTION

The following problem was raised by H.- J. Stoss [3] in connection with certain questions related to the complexity of Boolean functions. An acyclic directed graph G is said to have property $P(m,n)$ if for any set X of m vertices of G, there is a directed path of length n in G which does not intersect X. Let $f(m,n)$ denote the minimum number of edges a graph with property $P(m,n)$ can have. The problem is to estimate $f(m,n)$.

For the remainder of the paper, we shall restrict ourselves to the case $m = n$. We shall prove

(1) $$c_1 n \log n / \log \log n < f(n,n) < c_2 n \log n$$

(where $c_1, c_2, \ldots,$ will hereafter denote suitable positive constants). In fact, the graph we construct in order to

* P. Erdös and E. Szemerédi, Mathematical Institute of the Hungarian Academy of Sciences, Réaltanoda ut. 13-15 Budapest V, Hungary.

** R. L. Graham, Bell Laboratories, Department of Mathematics, 500 Mountain Avenue, Murray Hill, New Jersey.

This research was supported in part by National Science Foundation grant DCR72-03752 A02, by the Office of Naval Research contract NR 044-402, and by IBM Corporation. Reproduction in whole or in part is permitted for any purpose of the United States Government.

establish the upper bound on $f(n,n)$ in (1) will have just c_3n vertices. In this case the upper bound in (1) is essentially best possible since it will also be shown that for c_4 sufficiently large, every graph on c_4n vertices having property $\mathcal{P}(n,n)$ must have at least $c_5 n \log n$ edges.

A PRELIMINARY LEMMA

In order to establish the upper bound in (1) we first need the following result.

Lemma. For all $\delta > 0$ there exists $c = c(\delta)$ such that for all t sufficiently large, there exists a bipartite graph $B = B(\delta;t)$ with vertex sets A and A' so that:

(i) $|A| = |A'| = t$;

(ii) B has at most $c(\delta)t$ edges;

(iii) If $X \subseteq A$, $X' \subseteq A'$ with $|X| \geq \delta t$, $|X'| > \delta t$ then $(X,X') = \{\{x,x'\}: x \in X, x' \in X'\}$ contains an edge of B.

Proof: We use a simple probabilistic argument to show the existence of B. Form a bipartite graph $\bar{B}$ on the vertex sets A and A' with $|A| = |A'| = t$ by selecting for each $a \in A$ a random subset $\bar{B}(a) \subseteq A'$ of cardinality $d = d(\delta)$ (to be specified later). Call $\bar{B}$ "bad" if there exists $X \subseteq A$, $X' \subseteq A'$, with $|X| \geq \delta t$, $|X'| > \delta t$, so that (X,X') contains no edge of $\bar{B}$. For fixed X and X', the probability that $\bar{B}$ is bad because of these two subsets is at most

$$\binom{(1-\delta)t}{d}^{\delta t} \Big/ \binom{t}{d}^{\delta t} < \left(\frac{(1-\delta)t}{t-d}\right)^{d\delta t}.$$

Hence, the total probability that $\bar{B}$ is bad is at most

$$\binom{t}{\delta t}^2 \left(\frac{(1-\delta)t}{t-d}\right)^{d\delta t} < 2^{2t}\left(\frac{1-\delta}{1-d/t}\right)^{d\delta t} .$$

A simple computation shows that if d is chosen suitably large, for example, so that

$$(1-\delta^2)^{d\delta} < 1/4,$$

then for t sufficiently large this probability is less than 1, and so, a graph $B = B(\delta;t)$ must exist which satisfies the requirements of the lemma. ∎

CONSTRUCTION OF G

The next step in the proof of (1) is the construction of the directed graph G. For large n, $G = G(n)$ will have as its vertex set the set $V = \{0,1,\ldots,2^n-1\}$. If v and m are positive integers, then $D_v(m)$ will denote the set $\{v, v+1, \ldots, v+m-1\} \cap V$. Similarly, $D_v^*(m)$ will denote the set $\{v, v-1, \ldots, v-m+1\} \cap V$. In general, $\varepsilon_1, \varepsilon_2, \ldots,$ will denote suitably chosen fixed positive constants to be specified later. The edge set E of G is formed as follows:

(i) For $v \in V$, the pairs (v,x), $x \in D_{v+1}(4n)$, are in E;

(ii) For each t with $n/2 \le 2^t < 2^n$ and each i as specified below, a copy of $B(\varepsilon_i;2^t)$ is formed between the vertex sets $A = D_{m\cdot 2^t}(2^t)$ and $A' = D_{(m+i)\cdot 2^t}(2^t)$, $0 \le m < 2^{n-t}$, where $i = 1,2,\ldots,10$ (or if i cannot assume the value 10 because $(m+10)2^t > 2^n$, then it ranges from 1 to $2^{n-t}-m$). All edges are directed from x to y with $x < y$.

An elementary calculation shows that

$$|E| < c_6 n 2^n.$$

THE UPPER BOUND

Theorem 1. For a suitable $\varepsilon > 0$, $G(n)$ has property $\mathcal{P}(\varepsilon\cdot 2^n, \varepsilon\cdot 2^n)$ for all sufficiently large n.

Proof: The theorem will be proved by a sequence of claims. First we show that $G(n)$ shares with the graphs $B(\varepsilon;t)$ the following property.

Claim 1. If $m \geq 2n$ and $X \subset D_x(m)$, $X' \subset D_{x+m}(m)$ satisfy $|X| \geq \varepsilon_2 m$, $|X'| \geq \varepsilon_2 m$, then $[X,X'] = \{(x,x') : x\in X,\ x'\in X'\}$ contains an edge of $G(n)$.

Proof of Claim: Let $2^t \leq m/2 < 2^{t+1}$. Thus, $m/4 < 2^t$ so at most five of the intervals $D_{r\cdot 2^t}(2^t)$ intersect $D_x(m)$ and at most five of them intersect $D_{x+m}(m)$. Since $|X| \geq \varepsilon_2 m$ then some $D_{r\cdot 2^t}(2^t)$ and $D_{r'\cdot 2^t}(2^t)$ have

$$(3)\qquad |D_{r\cdot 2^t}(2^t) \cap X| \geq \varepsilon_2 m/5,\ |D_{r'\cdot 2^t}(2^t) \cap X'| \geq \varepsilon_2 m/5.$$

But we must have $|r'-r| \leq 10$ so that by the construction of $G(n)$ there is a copy of $B(\varepsilon_1;2^t)$ between $D_{r\cdot 2^t}(2^t)$ and $D_{r'\cdot 2^t}(2^t)$. Thus, if $\varepsilon_2/5 > \varepsilon_1$ and $m > 2^t$ then the property of $B(\varepsilon_1;2^t)$ guaranteed by the Lemma implies that $[X,X']$ contains an edge of $G(n)$ provided that t is sufficiently large (which is guaranteed by choosing n large enough). This proves the claim. ∎

Next, let us choose an arbitrary fixed set X of vertices with $|X| \leq \varepsilon \cdot 2^n$. The vertices in X will be referred to as the marked vertices of G; the remaining vertices of G will be called the unmarked vertices of G.

Let us call an unmarked vertex $y \in V$ bad if for some $m \geq 1$ either at least $\varepsilon_3 m$ vertices in $D_y(m)$ are marked or at least $\varepsilon_3 m$ vertices in $D_y^*(m)$ are marked. Otherwise, an unmarked vertex of G is called good.

Claim 2. There are at most $\varepsilon_4 2^n$ bad vertices.

Proof of Claim: Let y_1 denote the least unmarked vertex of G (if it exists) for which for some $m_1 \geq 1$, at least $\varepsilon_3 m_1$ vertices in $D_{y_1}(m_1)$ are marked. In general, if $y_1, \ldots, y_k$ and $m_1, \ldots, m_k$ have been defined, let y_{k+1} be the least unmarked vertex of G following $y_k + m_k - 1$ (if it exists) for which for some $m_{k+1} \geq 1$ at least $\varepsilon_3 m_{k+1}$ vertices in $D_{y_{k+1}}(m_{k+1})$ are marked. We continue this process until it no longer can be applied, so that, say, $y_1, \ldots, y_s$ and $m_1, \ldots, m_s$ have been defined. Similarly, let y_1^* denote the greatest unmarked vertex (if it exists) for which for some $m_1^* \geq 1$, at least $\varepsilon_3 m_1^*$ vertices in $D_{y_1^*}^*(m_1^*)$ are marked, etc. In this way, we define $y_1^*, \ldots, y_{s^*}^*$ and $m_1^*, \ldots, m_{s^*}^*$.

It follows from the preceding construction and the definition of a bad vertex that all bad vertices are contained in the set

$$Y = \bigcup_{k=1}^{s} D_{y_k}(m_k) \cup \bigcup_{k=1}^{s^*} D^*_{y^*_k}(m^*_k)$$

Thus, there are at most

$$M = \sum_{k=1}^{s} m_k + \sum_{k=1}^{s^*} m^*_k$$

bad vertices. However, by our construction there are at least $(\varepsilon_3/2)M$ marked vertices in Y. Since by hypothesis there are at most $\varepsilon \cdot 2^n$ marked vertices in V then we have

$$(\varepsilon_3/2)M \leq \varepsilon \cdot 2^n,$$

$$M \leq (2\varepsilon/\varepsilon_3)2^n < \varepsilon_4 2^n,$$

which proves the claim. ■

For an unmarked vertex x, let $\underline{P_x(m)}$ denote the set of all unmarked vertices in $D_x(m)$ which can be reached from x by directed paths which contain only unmarked vertices.

Claim 3. If x is a good vertex and $|D_x(m)| = m$ then

(4) $$|P_x(m)| > \varepsilon_5 m$$

Proof of Claim: If $m \leq 4n$ then since x is good, at least $(1-\varepsilon_3)m$ vertices in $D_x(m)$ are unmarked and x has edges directly to all of them. Suppose $m > 4n$. Let m' denote $[m/2]$. Since $|D_x(m')| = m'$ then by induction $|P_x(m')| > \varepsilon_5 m'$. Since x is good then at most $\varepsilon_3 m$ vertices in $D_x(m)$ are marked. Hence, at most $\varepsilon_3 m$ vertices in $D_{x+m'}(m') \subseteq D,(m)$ are marked. Since $m' \geq 2n$ and $\varepsilon_5 \geq \varepsilon_2$ then there are edges from $P_x(m')$ to at least $(1-\varepsilon_2)m'$ vertices in $D_{x+m'}(m')$. But at most $\varepsilon_3 m < 3\varepsilon_3 m'$ vertices in $D_{x+m'}(m')$ are marked. Hence, $P_x(m')$ must have edges to at least $(1-\varepsilon_2-3\varepsilon_3)m'$ unmarked vertices in $D_{x+m'}(m')$. Since $1-\varepsilon_2-3\varepsilon_3 > 3\varepsilon_5$ then

$$|P_x(m)| > 3\varepsilon_5 m' > \varepsilon_5 m.$$

The claim now follows by induction, $\square$

In exactly the same way it follows that if $P_x^*(m)$ denotes the set of all unmarked vertices in $D:(m)$ which are connected to the unmarked vertex x by a directed path containing only unmarked vertices, and x is a good vertex and $|D_x^*(m)| = m$, then-

$$|P_x^*(m)| > \varepsilon_5 m. \tag{4'}$$

Claim 4. Let x -and x' be good vertices with $x < x'$. Then $x' \in P_x(2^n)$.

Proof : If $x'-x \leq 4n$ then the claim is immediate since by construction there is an edge from x to x'. Assume $x'-x > 4n$. Let $y = [(x+x')/2]$ and let $m = y - x + 1$.

Consider the intervals $D_x(m)$ and $D^*_{x'}(m)$. Either they are adjacent or they have the single element y in common. Since x and x′ are good then by (4) and (4′)

$$(5) \qquad |P_x(m)| > \varepsilon_5 m, \; |P^*_{x'}(m)| > \varepsilon_5 m.$$

Since $\varepsilon_5 > \varepsilon_2$ then by Claim 1, there is an edge in G from a vertex of $P_x(m)$ to a vertex of $P^*_{x'}(m)$. Thus, there is a directed path from x to x′ containing no marked vertices and the claim is proved. ■

The proof of the theorem is now immediate. By Claim 2 there are at least $(1-\varepsilon_4-\varepsilon)2^n$ good vertices in G. By Claim 4 we can form a directed path which contains only unmarked vertices and which contains <u>all</u> the good vertices (since x′ can always be chosen to be the next good vertex following x). Since $1-\varepsilon_4-\varepsilon > \varepsilon$ then the theorem follows (where it is easily seen how the appropriate values of ε_k and c_k can be chosen). ■

<u>THE LOWER BOUND</u>

The following result will establish the lower bound in (1).

<u>Theorem 2</u>. Let H be an acyclic directed graph with at most $c_7 n \log n/\log\log n$ edges where n is a large fixed integer. Then there is a set of at most n vertices of H which hits every directed path of length n.

<u>Proof</u>: Let us denote the vertex set of H by $V = \{1,2,\ldots,v\}$. We may assume that H has at least $c_8\, n \log n/\log\log n$ edges. We may also assume that all edges are of the form (i,j) with

$i < j$. For an edge $e = (i,j)$ of H, let length (e) be defined to be $j-i$. Partition the edges of H into classes $C_0, C_1, \ldots, C_r$ where

$$C_k = \{e : 2^{4k \log\log n} < \text{length}(e) < 2^{4(k+1)\log\log n}\}$$

and $r = [\log v/4 \log\log n]$.

Since H has at least $c_8\, n \log n/\log\log n$ edges then it follows that $v \geq c_9\, n^{1/2}$ and $r \geq c_{10} \log n/\log\log n$. Hence some class C_a with $0 \leq a < r$ has at most $c_{11}\, n$ elements, Let us delete all vertices in H incident to any of the edges in C_a. Furthermore, we also delete those vertices $x \in V$ which satisfy

$$0 \leq x - m \cdot 2^{4a \log\log n}(1 + 2^{2 \log\log n}) < 2^{4a \log\log n}$$

for some integer $m \geq 0$. This latter step removes at most

$$\left(\frac{2}{2^{2 \log\log n} - 1}\right) v = o(n)$$

vertices, since $v \leq 2\, c_7\, n \log n/\log\log n$. Hence we have deleted at most $c_{12}\, n$ vertices altogether. However, any directed path remaining has at most

$$\left(\frac{2^{(4a+2)\log\log n} - 2^{4a \log\log n}}{2^{4(a+1)\log\log n}}\right) v = o(n)$$

edges, since we cannot go more than $2^{(4a+2)} \log\log n - 2^{4a} \log\log n$ steps without using an edge whose length exceeds $2^{4a} \log\log n$; and the length of such an edge actually exceeds $2^{4(a+1)} \log\log n$. This proves the theorem. ∎

By using a different partition of the edges of H, namely, into the classes $C_0', \ldots, C_r'$, where

$$C_k' = \{e : 2^{c_{13}k} \leq \text{length}(e) < 2^{c_{13}(k+1)}\}$$

for a suitable constant c_{13}, we can establish the following result.

<u>Theorem 3</u>. If c_{14} is sufficiently large then any graph G on $c_{14}n$ vertices having property $\mathcal{P}(n,n)$ must have at least $c_{15}n \log n$ edges.

The graphs G(n) used in Theorem 1 show that the result in Theorem 3 is best possible to within constant factors.

SOME RELATED QUESTIONS

We now consider several problems for ordinary (undirected) graphs. Let $F_e(n,n)$ (resp., $F_v(n,n)$) denote the smallest integer for which there is a graph with $F_e(n,n)$ (resp., $F_v(n,n)$) edges so that with the deletion of any n of its vertices there still remains a connected component of n edges (resp., vertices). We shall prove by probabilistic methods that

(6) $$F_e(n, n) < c_{16}n, \quad F_v(n, n) < c_{17}n.$$

The method we use is the same as that in the work of Erdös and Rényi [1], [2]. It turns out that almost all graphs have the desired property.

Theorem 4. For every $\varepsilon > 0$ there is a $c = c(\varepsilon)$ so that all but $o\left(\binom{\binom{(2+\varepsilon)n}{2}}{cn}\right)$ graphs G with $(2+\varepsilon)n$ vertices and cn edges have the property that after the omission of any n of its vertices, a connected component of at least n vertices remains.

Proof: It suffices to show that if n vertices are omitted and the remaining $n(1+\varepsilon)$ vertices are split into two classes S_1 and S_2 with $|S_1| > \varepsilon n$, $|S_2| \geq \varepsilon n$, then there is at least one edge joining a vertex of S_1 to a vertex of S_2.

Consider a random graph G on $(2+\varepsilon)n$ vertices and cn edges (where c will be specified later). There are $\binom{(2+\varepsilon)n}{n}$ ways that n vertices of G can be deleted. The remaining $n(1+\varepsilon)$ points can then be split into two sets S_1 and S_2 in at most $2^{n(1+\varepsilon)}$ ways. Thus, the total number of splittings is at most

$$\binom{(2+\varepsilon)n}{n} 2^{n(1+\varepsilon)} < 2^{(2+\varepsilon)n} 2^{n(1+\varepsilon)} < 2^{3(1+\varepsilon)n}.$$

Between S_1 and S_2 there are at least εn^2 potential edges. The probability that none of these edges actually occurs in G is less than $\left(1 - \frac{c}{(2+\varepsilon)n}\right)^{\varepsilon n^2}$. Thus, if c is chosen so that

$$(7) \qquad 2^{3(1+\varepsilon)n}\left(1 - \frac{c}{(2+\varepsilon)n}\right)^{\varepsilon n^2} \to 0$$

as $n \to \infty$ then we easily see that almost all graphs cannot be split in such a way.

Since

$$\left(1 - \frac{c}{(2+\varepsilon)n}\right)^{\varepsilon n^2} \to e^{-(\frac{\varepsilon c}{2+\varepsilon})n}$$

then for c large enough, e.g., $c > 18(\varepsilon+\varepsilon^{-1})$,

$$e^{-(\frac{\varepsilon c}{2+\varepsilon})n} < e^{-3(1+\varepsilon)n}$$

and (7) holds. This proves the theorem. ∎

The other half of (6) is proved in a similar way. It would be interesting to determine the best possible value of c but this does not seem to be too easy.

We mention here the undirected analogue of (1). Let $g(n,n)$ denote the smallest integer for which there is an undirected graph of $g(n,n)$ edges so that if we omit any n of its vertices then there always remains a path of length n. We believe

$$\frac{g(n,n)}{n} \to \infty, \quad \frac{g(n,n)}{n \log n} \to 0$$

as $n \to \infty$ and hope to return to this question in finite time.

A related question is the following: Consider random graphs on n vertices and Cn edges. Is it true that for large C almost all of these graphs have a path of length $n(1-\varepsilon)$? It is known [4] that almost all graphs on n vertices and $(\frac{1}{2} + \varepsilon)$ n log n edges are Hamiltonian.

It is possible to introduce another parameter into these questions. Let $F_v(t;n,n)$ denote the smallest integer for which there is a graph with t vertices and $F_v(t;n,n)$ edges having the property that if any n vertices are deleted there still remains a connected component with at least n vertices. If $t/n \to c > 2$ then $F_v(t;n,n)/n \to A(c)$ where $A(c) \to \infty$ as $c \to 2$. (The behavior of $F_e(t;n,n)/n$ is similar). We could also omit edges instead of vertices but leave the formulation of these questions to the reader.

Acknowledgment.

The authors gratefully acknowledge the useful suggestions on this problem given to us by Don Knuth.

REFERENCES

1. P. Erdös and A. Rényi, On random graphs I, Publ. Math. Debrecen 6 (1959) 290-297.

2. P. Erdös and A. Rényi, On the evolution of random graphs, Bull. Int. Stat. Inst. 38 (1961) 343-347.

3. D. E. Knuth (personal communication)

4. J. Komlós and E. Szemerédi, On Hamiltonian circuits in random graphs, (to appear).