Large Induced Trees in Sparse Random Graphs

A. M. FRIEZE

*Department of Computer Science and Statistics, Queen Mary College,
London University, London, England*

AND

B. JACKSON

Department of Mathematics, Goldsmiths College, London University, London England

We consider the size of the largest induced tree in random graphs, random regular graphs and random regular digraphs where the average degree is constant. In all cases we show that with probability $1 - o(1)$, such graphs have induced trees of size order n. In particular, the first result confirms a conjecture of Erdos and Palka (*Discrete Math.* **46** (1983), 145–150).   © 1987 Academic Press, Inc.

1. INTRODUCTION

This paper is concerned with the order of the largest induced tree in various models of random graphs (an induced tree being a vertex induced subgraph which is a tree). Previous research has concentrated on the random graph $G_{n,p}$ which has vertices $V_n = \{1, 2, ..., n\}$ and each of the $N = \binom{n}{2}$ possible edges are included independently with probability $p = p(n)$ and excluded with probability $q = 1 - p$.

For a graph G let $\tau(G)$ denote the order of its largest induced tree. Erdös and Palka [6] showed that if p is constant then

$$\frac{\tau(G_{n,p})}{\log n} \to \frac{2}{\log(1/q)}$$

with probability 1 as $n \to \infty$. See also Marchetti-Spaccemela and Protasi [10] and Palka and Rucinski [11].

It was conjectured in [6] that if $p = c/n$, c constant, $c > 1$, then there exists $\phi(c) > 0$, independent of n, such that

$$\tau(G_{n,p}) \geq \phi(c) n \qquad \text{a.s.}$$

181

(A property π_n will be held to hold *almost surely* (a.s.) if $\lim_{n \to \infty} \Pr(\pi_n) = 1$). We give an outline proof of this conjecture plus detailed proofs of two related results. We prove these by analysing the performance of a simple algorithm. Our current estimate for the largest possible value of $\phi(c)$ is rather weak. We state what we have proved as

THEOREM 1.1.   *If* $p = c/n$, *c constant*, $c > 1$, *then*

$$\tau(G_{n,p}) \geqslant c^{-2} \min\{\sqrt{c-1}, \tfrac{1}{2}\} n \qquad a.s.$$

One of the referees pointed out that Fernandes-de-la-Vega [7] has independently proved this conjecture. He has shown that a.s. $\tau(G_{n,p}) \geqslant \alpha_c n$ where α_c is the least positive root of $cx = \log(1 + c^2 x)$. This gives a stronger result than Theorem 1.1 as stated. For this reason we only give an outline of our proof of Theorem 1.1. (The reader may also be interested to learn that in [8] we have been able to show that $G_{n,p}$ a.s. contains an induced cycle of length $\beta(c) n$ when $p = c/n$, c a large enough constant, and $\beta(c) > 0$ independent of n.)

We can also analyze the performance of or algorithm on other "sparse" random graphs. In particular we consider random regular graphs and digraphs.

Let $\mathscr{R}(r, n)$ denote the set of r-regular graphs with vertex set V_n. We turn $\mathscr{R}(r, n)$ into a probability space by giving each graph the probability $1/|\mathscr{R}(r, n)|$. Let $RG(r, n)$ denote a random graph chosen from $\mathscr{R}(r, n)$. Our second result is

THEOREM 1.2.   *Let* $r \geqslant 3$ *be constant. Then*

$$\tau(RG(r, n)) \geqslant (1 - o(1)) \frac{rn}{2(r-1)} \left(1 - (r-2) \log_e \frac{r-1}{r-2}\right) \qquad a.s.$$

(*the* $o(1)$ *term tends to zero as n tends to* ∞, *naturally*).

We shall also consider random digraphs of constant outdegree. Thus let $\mathscr{D}(r, n)$ denote the set of digraphs of regular outdegree r and vertex set V_n. We turn $\mathscr{D}(r, n)$ into a probability space by giving each digraph the probability $1/|\mathscr{D}(r, n)|$. Let $D(r, n)$ denote a random graph chosen from $\mathscr{D}(r, n)$. Our final result is

THEOREM 1.3.   *Let* $r \geqslant 2$ *be constant. Then*

$$\tau(D(r, n)) \geqslant (1 - o(1)) \frac{rr^*}{6(r-1)} \left(3 - \frac{r^3 r^{*2}}{r+1}\right) n \qquad a.s.$$

where r^* is the smallest root of

$$r^2 x = (1 - r^2 x)\left(r - \frac{r^2 x}{r+1}\right).$$

(Note that $r^* < r^{-2}$.)

Note that the tree found in Theorem 1.3 is an arborescence. We shall prove our theorems in the order 1.2, 1.3, 1.1 (outline only).

2. THE BASIC ALGORITHM

We describe an algorithm *TF* for finding an induced tree in a graph $G = (V_n, E)$. At a general stage we will have a set of vertices T and a subset $L \subseteq T$. The set $T - L$ will be known to induce a tree in G. The vertices in L are *unprocessed* leaves of a larger tree defined on T. This larger tree may not be vertex induced. This lack of knowledge comes from not having yet explored all the edges incident with vertices in L.

In order to avoid some clumsy statements about the initial position, we add a vertex 0 and the edge $\{0, 1\}$ and start with $T = \{0, 1\}$ and $L = \{1\}$. The vertices $V_n - T$ are partitioned into 2 sets, F and B. The set of *free* vertices F have not yet been encountered by *TF* and the set of *bad* vertices B have been permanently excluded from T. Initially $F = \{2, 3, ..., n\}$ and $B = \varnothing$.

The general step of the algorithm is to choose $v \in L$ and examine its incident edges using the procedure PROCESS(v). Let $E(v) = \{e \in E: v \in e\} = \{e_i = \{v, v_i\}: i = 1, 2, ..., d(v)\}$, where $d(v)$ is the degree of v in G. Because we are using the model of Bollobás [1] to examine regular graphs we will need to allow the existence of loops and multiple edges.

By construction there will be a unique $w \in T - L$ such that $\{v, w\} \in E$. Let w be denoted by $p(v)$ and assume that $v_1 = p(v)$. PROCESS(v) examines the edges $e_2, e_3, ..., e_d(v)$. We use 2 versions of PROCESS:

OPTION 1 (*Regular graphs only*)
for $i = 2$ *to* $d(v)$ *do*
begin
    if $v_i \in F$ *then* $T := T + v_i;\ L := L + v_i;\ F := F - v_i;\ p(v_i) := v$
    else if $v_i \in L$ *then* $T := T - v_i;\ L := L - v_i;\ B := B + v;$
    {Remark: observe that we do nothing if $v_i \in B + v$.}
end;
$L := L - v$.
(Notation: $X + a = X \cup \{a\},\ X - a = X - \{a\}$)

OPTION 2 (*Regular digraphs and* $G_{n,p}$)
if $L \cap \{v_2, v_3, ..., v_{d(v)}\} \neq \varnothing$ then $T := T - v$; $L := L - v$; $B := B + v$
else for $i = 2$ to $d(v)$ do
$\quad$ if $v_i \in F$ then begin $T := T + v_i$; $L := L + v_i$; $F := F - v_i$; $p(v_i) := v$ end;

$L := L - v$.

The 2 options differ in which vertex is deleted when we discover an edge $\{v, v_i\}$ where $v_i \in L$.

ALGORITHM *TF.*
begin
$\quad$ $T := \{0, 1\}$; $L := \{1\}$; $F := \{2, 3, ..., n\}$; $B := \varnothing$; $p(1) := 0$;
$\quad$ *while* $L \neq \varnothing$ *do*
$\quad$ *begin*
$\quad\quad$ choose $v \in L$; PROCESS(v)
$\quad$ *end*
end

The following lemma is easy to prove, but it is important that we check its truth.

LEMMA 2.1. *Prior to each execution of* PROCESS

$\quad$ (a) $(T - L) \cap \{v_2, v_3, ..., v_{d(v)}\} = \varnothing$.
$\quad$ (b) $H = (T, E_T)$ *is a tree, where* $E_T = \{(v, p(v)) : v \in T - 0\}$.
$\quad$ (c) $T - (L + 0)$ *is an* induced *tree of G.*
$\quad$ (d) $v \in L$ *implies v is a leaf of H.*

Proof. (a) holds because if $w = v_i \in T - L$ then either

$\quad$ $w \neq p(v)$: PROCESS(w) would have put v into B—contradiction

or

$\quad$ $w = p(v)$: PROCESS(w) would first have put v into L and then into B
$\quad\quad\quad\quad\quad\quad$ when a second edge $\{v, w\}$ was found—contradiction.

(Note that $w \in T - L \rightarrow$ PROCESS(w) has already een executed.) We have only to check that the truth of the remaining statements is unaffected by an execution of PROCESS.

OPTION 1. We consider the possibilities for v_i. It should be clear that (b) – (d) continue to hold in all cases:

$v_i \in F$: H grows by the addition of leaf $v_i \in L$. $T - L$ is unchanged.

$v_i \in B + v$: nothing changes.

$v_i \in L$: since v_i is a leaf of H we can just delete it. $T - L$ is unchanged.

At the end of PROCESS(v) we put v into $T - L$. At this stage v is a leaf of the tree induced by $T - L$ because it was a leaf of H before execution of process and no new vertices have been added to $T - L$ other than v.

OPTION 2. If $L \cap \{v_2, v_3, ..., v_{d(v)}\} \neq \emptyset$ we just delete v from H. This is justified because v is a leaf of H at this stage. Otherwise we just add the neighbors of v in F to L as leaves of H. Putting v into $T - L$ at the end does not affect (b)–(d) by the same argument as in OPTION 1.

3. RANDOM REGULAR GRAPHS

Let δ, Δ be integer *constants* satisfying $3 \leqslant \delta \leqslant \Delta$. Let $\mathbf{d} = d_1 d_2 \cdots d_n$ satisfy $\delta \leqslant d_i \leqslant \Delta$ for $i = 1, 2, ..., n$.

Let $\mathscr{G}(\mathbf{d})$ denote the set of simple graphs with vertex set $V_n = \{1, 2, ..., n\}$ satisfying $d(i) = d_i$. Thus graphs in $\mathscr{G}(\mathbf{d})$ have no loops or multiple edges. Assume that $\mathbf{d}$ is graphic, i.e., $\mathscr{G}(\mathbf{d}) \neq \emptyset$. We turn $\mathscr{G}(\mathbf{d})$ into a probability space by giving all members of $\mathscr{G}(\mathbf{d})$ the same probability $1/|\mathscr{G}(\mathbf{d})|$. Theorem 1.2 follows from Lemma 3.1 when we take $\delta = \Delta = r$.

LEMMA 3.1. *Let G be chosen at random from $\mathscr{G}(\mathbf{d})$. Then*

$$\lim_{n \to \infty} \Pr(\tau(G)) \geqslant (1 - o(1)) \frac{nd}{2(\Delta - 1)} \left(1 - (\Delta - 2) \log_e \frac{\Delta - 1}{\Delta - 2}\right) = 1$$

where $nd = d_1 + d_2 + \cdots + d_n$.

Proof. In order to study $\mathscr{G}(\mathbf{d})$ we consider the model defined in Bollobás [1]. Let $D_1, D_2, ..., D_n$ be disjoint sets with $|D_i| = d_i$ and set

$$D = \bigcup_{i=1}^{n} D_i \quad \text{and} \quad 2a = |D| = nd.$$

A *configuration* C is a partition of D into a pairs, the *edges* of C. Let Φ be the set of all $N(a) = (2a)! \, 2^{-a}/a!$ configurations. Turn Φ into a probability space by giving all members of Φ the same probability. For $C \in \Phi$ let $\phi(C)$ be the *multi-graph* with vertex set V_n in which i is joined to j whenever C has an edge with one end-vertex in D_i and the other in D_j. Clearly $\mathscr{G}(\mathbf{d}) \subseteq \phi(\Phi)$ and

$$|\phi^{-1}(G)| = \prod_{i=1}^{n} d_i!$$

for every $G \in \mathscr{G}(\mathbf{d})$.

Let Q be a property of the graphs in $\mathscr{G}(\mathbf{d})$ and let Q^* be a property of the configurations in Φ. Suppose these properties are such that for $G \in \mathscr{G}(\mathbf{d})$ and $C \in \phi^{-1}(G)$ the configuration C has Q^* if and ony if G has Q. All we shall need from [1] is that if almost every C has Q^* then almost every G has Q.

We shall thus be able to prove the lemma is we can show that TF applied to a multigraph $\phi(C)$, C chosen randomly from Φ, almost surely finds an induced tree of the given size.

The tree produced by TF can contain loops, but will not have multiple edges. This is no problem, because then by the above, TF a.s. works well on $G(\mathbf{d})$ where there are no loops.

It is important to realize that we can sample uniformly from Φ as follows: Let x_1 be chosen in any way from D and y_1 be chosen uniformly at random from $D - \{x_1\}$. Having chosen pairs $\{x_1, y_1\}$, $\{x_2, y_2\}, ...,$ $\{x_k, y_k\}$, let $X = D - \{x_1, y_1, x_2, x_y, ..., x_k, y_k\}$. Let x_{k+1} be chosen in any way from x and then let y_{k+1} be chosen uniformly at random from $X - \{x_{k+1}\}$. Repeat the whole process a times to produce C.

To prove the lemma we shall apply algorithm TF and construct C in tandem. At the general point in the algorithm let $D'_v = \{x \in D_v : x$ has not yet been paired in the construction of $C\}$ and for $X \subseteq V_n$ let $D'_X = \sum_{v \in X} |D'_v|$. We shall now represent TF as algorithm $TF1$. We will add statements updating certain parameters t, f, b, l, m. The meanings of these parameters are

$$t = |T|, \quad f = D'_F, \quad b = D'_B, \quad l = D'_L, \quad \text{and} \quad m \text{ is an iteration count.} \quad (3.1)$$

ALGORITHM $TF1$.

begin
    $T := \{0, 1\}$; $L := \{1\}$; $F := \{2, 3, ..., n\}$; $B := \varnothing$;
    $t := 1$; $l := d(1)$; $f := d(2) + d(3) + \cdots + d(n)$; $b := 0$;
    $p(1) := 0$; $m := 0$;
    while $L \neq \varnothing$ *do*
        choose $v \in L$; PROCESS(v)
    end.
procedure PROCESS(v)
begin
    let $E(v) = \{\{v, v_i\} : i = 1, 2, ..., d(v)\}$ where $v_1 = p(v)$;
    assume that if there are λ loops at vertex v then $v_i = v$ for $i \geqslant d(v) -$
    $2\lambda + 1$;
    $k := d(v) - 1$;
    $i := 1$;

LOOP: *repeat*
  $m := m + 1; \; i := i + 1;$
  case of
 (a) $v_i \in F$: $T := T + v_i; \; L := L + v_i; \; F := F - v_i;$
       $t := t + 1; \; l := l + d(v_i) - 2; \; f := f - d(v_i);$
       $p(v_i) := v; \; k := k - 1.$
 (b) $v_i \in L - \{v\}$: $T := T - v_i; \; := L - v_i; \; B := B + v_i;$
       $t := t - 1; \; l := l - d(v_i); \; b := b + d(v_i) - 2;$
       $k := k - 1.$
 (c) $v_i \in B$: $l := l - 1; \; b := b - 1; \; k := k - 1.$
 (d) $v_i = v$: $l := l - 2; \; k := k - 2; \; i := i + 1.$
  end
  until $k = 0;$
 $L := L - \{v\}$
end.

When we come to execute PROCESS(v) we will find that one element x of D_v has been paired with an element of $D_{p(v)}$ but the elements in $D_v - \{x\}$ are as yet unpaired. In an arbitrary execution of the LOOP there will be k elements of D_v which have not been paired. We choose one, say y. We then randomly choose z from all the unpaired elements of D and pair it with y to make an edge of C and hence $\phi(C)$. Suppose $z \in D_w$. We then execute cases (a), (b), (c), (d) according as $w \in F, L - \{v\}, B, \{v\}$.

We claim (see (3.1)) and the reader can easily verify by induction on m that the choices for z divide into

f choices such that $w \in F$.

$l - 1$ choices such that $w \in L, k - 1$ of which are

such that $w = v$.

b choices such that $w \in B$.

We first discuss the probability that TF halts quickly with $L = \varnothing$, which is equivalent to $l = 0$.

Let E_m denote the event: case (a) occurs at the mth execution of LOOP. Let X_m be the random variable defined by

$$X_m = \begin{cases} \delta - 2 & \text{if } E_m \text{ occurs,} \\ -\delta & \text{otherwise.} \end{cases}$$

We claim that

$$\text{when } TF \text{ halts,} \quad X_1 + X_2 + \cdots + X_m \leqslant 0. \tag{3.2}$$

To see this let α, β, γ, λ denote the number of executions of cases (a), (b), (c), (d) before termination. Let $P = \{v: v = v_i \text{ in some execution of case (a)}\}$ and $Q = \{v: v = v_i \text{ in some execution of case (b)}\}$. Note that $Q \subseteq P$. Since $l = 0$ on termination

$$\gamma + 2\lambda + \sum_{v \in Q} d(v) = \sum_{v \in P} d(v) - 2\alpha.$$

Thus

$$\gamma + 2\lambda + 2\alpha = \sum_{v \in P \cdot Q} d(v) \geqslant \delta(\alpha - \beta).$$

Hence

$$\delta(\beta + \gamma + \lambda) \geqslant (\delta - 2)\,\alpha$$

and (3.2) follows. Now

$$\Pr(E_m) = f/(f + l + b - 1).$$

It is easy to check, by induction on m, that $b + l \leqslant (m + 1)(\varDelta - 2) + 2$ and $f \geqslant nd - (m + 1)\varDelta$ throughout. Thus

$$\Pr(E_m) \geqslant \frac{nd - (m + 1)\varDelta}{nd - 2m - 1} \qquad \text{regardless of the history of the algorithm up to this point.} \quad (3.3)$$

Thus if $\{Y_m\}$ is a set of independent random variables satisfying

$$\Pr(Y_m = \delta - 2) = \frac{nd - (m + 1)\varDelta}{nd - 2m - 1},$$

$$\Pr(Y_m = -\delta) = \frac{(\varDelta - 2)\,m + 1}{nd - 2m - 1},$$

then

$$\Pr(Y_1 + Y_2 + \cdots + Y_m \leqslant 0) \geqslant \Pr(X_1 + X_2 + \cdots + X_m \leqslant 0) \qquad \text{for all } m. \tag{3.4}$$

We note first that

$$\Pr(\exists m: 1 \leqslant m \leqslant n^{1/3} \quad \text{and} \quad Y_m = -\delta) \leqslant \frac{(\varDelta - 2)\,n^{2/3} + n^{1/3}}{nd - 2n^{1/3} + 1} \tag{3.5}$$

$$= o(1).$$

Now let

$$u_m = E(Y_m) = \frac{(nd - m\Delta)(\delta - 2) - \delta(\Delta - 2)m}{nd - 2m}(1 + o(1)), \qquad m \leqslant nd/3,$$

and

$$U_m = \sum_{i=1}^{m} u_m.$$

It follows from Theorem 2 of Höeffding [9] (stated as Lemma 3.2 below) that

$$\Pr(Y_1 + Y_2 + \cdots + Y_m \leqslant (1 - \varepsilon) U_m) \leqslant e^{-\varepsilon^2 U_m^2 / 2m(\delta - 1)^2} \qquad (3.6)$$

for any ε, $0 < \varepsilon < 1$.

Now for m large, we can write

$$U_m = (1 + o(1)) \int_0^m \frac{(nd - x\Delta)(\delta - 2) - \delta(\Delta - 2)x}{nd - 2x} \, dx$$

$$= (1 + o(1)) \int_0^m \left(\frac{\Delta(\delta - 2) + \delta(\Delta - 2)}{2} - \frac{nd(\delta - 1)(\Delta - 2)}{nd - 2x} \right) dx$$

$$= (1 + o(1)) \left(\frac{\Delta(\delta - 2) + \delta(\Delta - 2)}{2} m - \frac{nd(\delta - 1)(\Delta - 2)}{2} \log_e\left(\frac{nd}{nd - 2m} \right) \right)$$

$$= (1 + o(1)) \psi(m) \qquad \text{say.} \qquad (3.7)$$

We shall see later on that we will need to have shown that

$$m \text{ almost surely reaches the value } m^* = \lceil nd/2(\Delta - 1) \rceil. \qquad (3.8)$$

This will follow from (3.4) and (3.6) if we can show that $\psi(m)^2/m$ is sufficiently large for $n^{1/3} < m \leqslant m^*$. We first note that

$$m \leqslant \theta nd \quad \text{implies} \quad \psi(m) \geqslant m\left(\delta - 2 - \frac{2\theta \delta\Delta}{1 - 2\theta} \right), \qquad \theta < \frac{1}{2}.$$

This follows on using the inequality $\log_e(1 + x) \leqslant x$. By choosing $\theta = \theta_0 = \frac{1}{2}(1 + 2\delta\Delta)$ we obtain

$$\psi(m) \geqslant m(\delta - 5/2) \qquad \text{if} \quad m \leqslant \theta_0 nd. \qquad (3.9)$$

Now $\psi'(m)$ decreases monotonically for $0 < m \leqslant nd/2$ and so ψ is concave in this range. We can thus concentrate on estimating $\psi(m^*)$,

$$\psi(m^*) = (1 + o(1)) \, nd \left(\frac{\varDelta(\delta - 2) + \delta(\varDelta - 2)}{4(\varDelta - 1)} - \frac{(\delta - 1)(\varDelta - 2)}{2} \log_e \left(\frac{\varDelta - 1}{\varDelta - 2} \right) \right)$$

$$= (1 + o(1)) \, \lambda(\delta, \varDelta) \, nd.$$

The reader can easily check that $\lambda(3, 3)$, $\lambda(3, 4)$, $\lambda(4, 4)$ are all strictly positive. We can thus assume $\varDelta \geqslant 5$. We then use $\log_e(1 + x) \leqslant x - x^2/2 + x^3/3$ for $0 < x < 1$ to obtain

$$4(\varDelta - 1) \, \lambda(\delta, \varDelta) \geqslant \varDelta(\delta - 2) + \delta(\varDelta - 2)$$

$$- 2(\delta - 1)(\varDelta - 1) \left(1 - \frac{1}{2(\varDelta - 2)} + \frac{1}{3(\varDelta - 2)^2} \right)$$

$$= -2 + 2(\delta - 1)(\varDelta - 1) \left(\frac{1}{2(\varDelta - 2)} - \frac{1}{3(\varDelta - 2)^2} \right)$$

$$\geqslant -2 + 2(\varDelta - 1) \left(\frac{1}{\varDelta - 2} - \frac{2}{3(\varDelta - 2)^2} \right) \quad \text{as} \quad \delta \geqslant 3$$

$$= 2(\varDelta - 4)/3(\varDelta - 2)^2.$$

Thus, using (3.9), we can write

$$U_m \geqslant \mu(\delta, \varDelta) \, m \qquad n^{1/3} < m \leqslant m^*.$$

where $\mu(\delta, \varDelta) > 0$. Then putting $\varepsilon = n^{-.1}$ in (3.6) yields (3.8).

Given (3.5) we concentrate on the growth of $t = |T|$ until $m = m^*$. We note that if E_m occurs then t increases by 1 and if E_m does not occur then t decreases by at most 1 on iteration m.

It follows from (3.3) that if t_m is the value of t at the end of iteration m then for any $a \in R$

$$\Pr(t_m \geqslant a) \geqslant \Pr(Z_1 + Z_2 + \cdots + Z_m \geqslant a) \qquad (3.10)$$

where $\{Z_m\}$ is a set of independent random variables with

$$\Pr(Z_m = 1) = \frac{nd - m\varDelta}{nd - 2m + 1}$$

$$\Pr(Z_m = -1) = \frac{(\varDelta - 2) m + 1}{nd - 2m + 1}.$$

Now let

$$v_m = E(Z_m) = \frac{nd - 2(\varDelta - 1) m - 1}{nd - 2m + 1}$$

and

$$V_m = v_1 + v_2 + \cdots + v_m.$$

It follows from Lemma 3.2 that

$$\Pr(Z_1 + Z_2 + \cdots + Z_m \leqslant (1 - \varepsilon) V_m) \leqslant e^{-\varepsilon^2 V_m^2 / 2m} \tag{3.11}$$

for any ε, $0 < \varepsilon < 1$.

Putting $m = m^*$ we see that t_{m^*} is almost surely at least

$$(1 - o(1)) V_{m^*} = (1 - o(1)) \int_0^{m^*} \frac{nd - 2(\varDelta - 1) x}{nd - 2x} dx$$

$$= (1 - o)) \int_0^{m^*} \left((\varDelta - 1) - \frac{nd(\varDelta - 2)}{nd - 2x} \right) dx$$

$$= (1 - o(1)) \left((\varDelta - 1) m^* - \frac{(\varDelta - 2) nd}{2} \log_e \frac{nd}{nd - 2m^*} \right)$$

$$= (1 - o(1)) \frac{nd}{2} \left(1 - (\varDelta - 2) \log_e \frac{\varDelta - 1}{\varDelta - 2} \right).$$

By considering H as a tree rooted at 1 with branching factor at most $\varDelta - 1$ at each node we see that H has at least $|T|/(\varDelta - 1)$ non-leaves. But all non-leaves are in $T - L$ and the result follows from Lemma 2.1(c).

LEMMA 3.2. *Let* $X_1, X_2, \dots, X_m$ *be independent random variables where* $a_i \leqslant X_i \leqslant b_i$ *for* $i = 1, 2, \dots, m$. *Let* $X = (X_1 + X_2 + \cdots + X_m)/m$ *and let* $\mu = E(X)$. *Then*

$$\Pr(X \leqslant \mu - t) \leqslant e^{-2m^2 t^2 / \sum_{i=1}^m (b_i - a_i)^2} \tag{3.12}$$

(In [9] the inequality is given for $X \geqslant \mu + t$, but (3.12) can be obtained by looking at $-X_i$, $i = 1, 2, \dots, m$).

4. RANDOM OUTREGULAR DIGRAPHS

We now analyze the performance of a modification of *TF* using OPTION 2, called *TF2*, applied to a random digraph D constructed as

follows: $\mathbf{d} = d_1, \, d_2, ..., d_n$ satisfies $2 \leqslant \delta \leqslant d_i \leqslant \Delta$, $i = 1, 2, ..., n$. Each $v \in V_n$ independently chooses a set X_v of d_v vertices at random from $V_n - v$. (The resulting digraph D has $\sum_{i=1}^{n} d_i$ arcs.

ALGORITHM *TF2*.

begin
    $T := \{1\}$; $L := \{1\}$; $F := \{2, 3, ..., n\}$; $B := \varnothing$;
    while $L \neq \varnothing$ *do*
    begin
        choose $v \in L$; PROCESS $((v)$
    end
end.

Procedure PROCESS(v);
begin
    let $X_v = \{v_1, v_2, ..., v_d\}$
    case of
    (a) $X_v \cap T \neq \varnothing$: $T := T - v$; $L := L - v$; $B := B + v$.
    (b) $X_v \cap T = \varnothing$: *for* $i = 1$ *to* d *do*
                 case of
                 (1) $v_i \in F$: $L := L + v_i$; $T := T + v_i$; $F := F - v_i$.
                 (11) $v_i \in B$: do nothing.
                 end;
                 $L := L - v$.
end

A slight modification of the argument of Lemma 2.1 yields

LEMMA 4.1. *Throughout the execution of TF2* $T - L$ *induces an arborescence in D and for each $w \in L$ there is a unique $v \in T$ such that (v, w) is an arc of D. Hence*

$$|T - L| \geqslant |T|/(\Delta - 1). \tag{4.1}$$

We will now prove the main result of this section. Theorem 1.3 will follow as a corollary when we take $\delta = \Delta = r$.

LEMMA 4.2. *Let D be chosen at random as above. Then*

$$\lim_{n \to \infty} \Pr(\tau(D)) \geqslant (1 - o(1)) \frac{n\delta\xi^*}{\delta(\Delta - 1)} \left(3 - \frac{\Delta^3\xi^{*2}}{\Delta + 1}\right) = 1$$

where ξ^* is the smaller root of

$$\xi\Delta^2 = (1 - \xi\Delta^2)\left(\delta - \frac{\xi\Delta\delta}{\Delta + 1}\right). \tag{4.2}$$

(Note that $\xi^* < \Delta^{-2}$).

Proof. Consider the mth execution of the procedure PROCESS. Let $t = |T|$, $l = |L|$, $f = |F|$, and $b = |B|$ at the start. We note that X_v will be chosen independently of T, L, F, B as we will not have examined any arcs leaving v. Hence we have

$$\Pr(\text{case (a) occurs}) \leqslant \frac{\Delta t}{n - 1} \leqslant \frac{m\Delta^2}{n - 1}.$$

The first inequality follows as $\Delta t/(n - 1)$ bounds the expected value of $|X_v \cap T|$. The second is obvious, as $t \leqslant m\Delta$.

We note next that

$$\Pr(\text{there are } k \text{ executions of case (b1) } | \text{ case (b)}) = \binom{f}{k}\binom{b}{d-k} \bigg/ \binom{b+f}{d}.$$

Hence

$$E(\text{number of executions of case (b1) } | \text{ case (b)})$$

$$= \frac{df}{b+f} \geqslant \delta\left(1 - \frac{\Delta m}{(\Delta + 1)n}\right) \quad \text{for} \quad m \leqslant n/\Delta.$$

The latter inequality follows from $d \geqslant \delta$, $b + f = n - t$, and $(\Delta + 1)b + t \leqslant \Delta m$, this being easily confirmable by induction. We want next to show that m a.s. reaches $m^* = \xi^* n$ where ξ^* is the smaller root of (4.2). Now if case (a) occurs then l decreases by 1, and if case (b) occurs then the expected increase in l is at last $\delta(1 - \Delta m/(\Delta + 1)n) - 1$. Thus let $X_1, X_2, \ldots, X_m$ be a sequence of independent random variables satisfying $-1 \leqslant X_i \leqslant \Delta$ and

$$u_i = E(X_i) = \frac{-m\Delta^2}{n - 1} + \left(1 - \frac{m\Delta^2}{n - 1}\right)\left(\delta\left(1 - \frac{\Delta m}{(\Delta + 1)n}\right) - 1\right).$$

The techniques used in § 3 show that provided $U_m = u_1 + u_2 + \cdots + u_m$ is sufficiently large for $m \leqslant m^*$ then m will a.s. reach m^*. (The case of small m is handled separately, as before.) Now

$$U_m = (1 + o(1))\int_0^m \left(-\frac{\Delta^2 x}{n} + \left(1 - \frac{\Delta^2 x}{n}\right)\left(\delta - 1 - \frac{\delta\Delta X}{(\Delta + 1)n}\right)\right) dx$$

$$= (1 + o(1))\left((\delta - 1)m - \frac{1}{2}m^2\left(\frac{\delta\Delta}{(\Delta + 1)m} + \frac{\delta\Delta^2}{n}\right) + \frac{\delta\Delta^3 m^3}{3(\Delta + 1)n^2}\right). \tag{4.3}$$

Now for $\xi \leqslant \xi^*$,

$$\xi \Delta^2 \leqslant (1 - \xi \Delta^2) \left(\delta - \frac{\xi \Delta \delta}{\Delta + 1} \right). \tag{4.4}$$

Putting $m = \xi n$ into (4.3) and using (4.4) yields

$$U_m \geqslant (1 - o(1)) \, n \left((\delta - 1) \, \xi - \frac{1}{2} \left(\delta \xi + \frac{\xi^3 \Delta^2 \delta}{\Delta + 1} - \xi^2 \Delta^2 \right) + \frac{\delta \Delta^3 \xi^3}{3(\Delta + 1)} \right)$$

$$= (1 - o(1)) \, n \left(\left(\frac{1}{2} \delta - 1 \right) \xi + \xi^2 \Delta^2 \left(3 - \frac{\delta \Delta \xi}{\Delta + 1} \right) \Big/ 6 \right)$$

$$\geqslant \varepsilon(\delta, \Delta) \, n$$

for some $\varepsilon(\delta, \Delta)$ since $\delta \geqslant 2$ and $\xi^* \leqslant 1/\Delta^2$.

Thus m a.s. reaches m^*. We now consider how large $|T|$ will be at this time. We note that if case (a) occurs then $|T|$ decreases by 1 and that if case (b) occurs then the expected increase in $|T|$ is at least $\delta(1 - \Delta m/(\Delta + 1) n)$. Thus the expected increase in $|T|$ at the mth iteration is at least

$$v_m = -\frac{m \Delta^2}{n - 1} + \left(1 - \frac{m \Delta^2}{n - 1} \right) \delta \left(1 - \frac{\Delta m}{(\Delta + 1) n} \right).$$

The lemma follows by computing $v_1 + v_2 + \cdots + v_{m^*}$ and simplifying with (4.2).

5. Sparse Random Graphs

We now come to the Erdös–Palka conjecture for $G_{n,p}$, $p = c/n$, $c > 1$ constant. We note that either of Theorems 1.2 or 1.3 can be used to prove this conjecture for large c. We simply construct a.s. a large subgraph of $G_{n,p}$ with the required distribution. For details as to how this can be done, see Bollobás [2] or Bollobás, Fenner, and Frieze [3]. The method discussed here will demonstrate the existence of a large induced tree for all $c > 1$.

We shall be using a modification of TF using OPTION 2. Let us first note one problem about applying the algorithm naively. There is a substantial probability that vertex 1 is in a component of $G_{n,p}$ of size $O(\log n)$!. We can get round this by applying TF again, using only vertices that have not been looked at before and so on. With probability tending to 1 we eventually grow a large tree. This is one way around the problem. We follow a path with a simpler analysis. We break the algorithm up into stages.

Stage 1. Let $\varepsilon > 0$ be fixed an small, and let $n_1 = \lceil \varepsilon n \rceil$. First consider the subgraph H_1 of $G_{n,p}$ induced by $\{n - n_1 + 1, n - n_1 + 2,..., n\}$. It has the same distribution as $G_{n_1,p}$. Since $n_1 p \approx \varepsilon c$ we know from Erdös and Renyi [4] that H_1 a.s. contains a component T_1 of size at least $a \log n$ which is a tree. Here $a = a(\varepsilon, c)$ is independent of n and its exact value is irrelevant to us.

Thus in Stage 1 we select the largest tree component T_1 of H_1.

Stage 2. We throw away the vertices in $H_1 - T_1$ and then apply algorithm *TF* starting with $L = \{v \in V_n : v > n_1$ and v is adjacent to exactly one vertex in $T_1\}$ and $T = T_1 \cup L$. The analysis is similar to before although there are some minor technical difficulties.

REFERENCES

1. B. BOLLOBÁS, A probabilistic proof of an asymptotic formula for the number of labelled regular graphs, *European J. Combin.* 1 (1980), 311–316.
2. B. BOLLOBÁS, Long pahts in sparse random graphs, *Combinatorica* 2 (1982), 223–228.
3. B. BOLLOBÁS, T. I. FENNER, AND A. M. FRIEZE, Long cycles in sparse random graphs, "Graph Theory and Combinatorics," (A volume in honour of Paul Erdös) (B. Bollobás, Ed.) pp. 59–64, Academic Press, London.
4. P. ERDÖS AND A. RENYI, On the evolution of random graphs, *Publ. Math. Inst. Hungar. Acad. Sci.* 5 (1960), 17–61.
5. P. ERDÖS AND J. SPENCER, "Probabilistic Methods in Combinatorics," Academic Press, New York, 1974.
6. P. ERDÖS AND Z. PALKA, Trees in random graphs, *Discrete Math.* 46 (1983), 145–150.
7. W. FERNADEZ-DE-LA-VEGA, "Induced Trees in Sparse Random Graphs," Research Report 214, CNRS, April 1985.
8. A. M. FRIEZE AND B. JACKSON, Large holes in sparse random graphs, submitted for publication.
9. W. HÖEFFDING, Probability inequalities for sums of bounded random variables, *J. Amer. Statist. Assoc.* 58 (1963), 13–30.
10. A. MARCHETTI-SPACCEMELA AND M. PROTASI, The largest tree in a random graph, *Theoret. Comput. Sci.* 23 (1983), 273–286.
11. Z. PALKA AND A. RUCINSKI, On the order of the largest induced tree in a random graph, to appear.