The Transitive Closure of a Random Digraph

Richard M. Karp †

ABSTRACT

In a random n-vertex digraph, each arc is present with probability p, independently of the presence or absence of other arcs. We investigate the structure of the strong components of a random digraph and present an algorithm for the construction of the transitive closure of a random digraph. We show that, when n is large and np is equal to a constant c greater than 1, it is very likely that all but one of the strong components are very small, and that the unique large strong component contains about $\Theta^2 n$ vertices, where Θ is the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$. Nearly all the vertices outside the large strong component lie in strong components of size 1. Provided that the expected degree of a vertex is bounded away from 1, our transitive closure algorithm runs in expected time $O(n)$. For all choices of n and p, the expected execution time of the algorithm is $O(w(n)\,(n \log n)^{4/3})$, where $w(n)$ is an arbitrary nondecreasing unbounded function. To circumvent the fact that the size of the transitive closure may be $\Omega(n^2)$ the algorithm presents the transitive closure in the compact form $(A \times B) \cup C$, where A and B are sets of vertices, and C is a set of arcs.

1. INTRODUCTION

The probability space of digraphs $\boldsymbol{D}_{n,\,p}$ is defined as follows: each point in the space is a digraph with vertex set $\{1,2,...,n\}$ having no loops or multiple arcs, and the probability of a given digraph D with e arcs is $p^e(1-p)^{n(n-1)-e}$. In other words, each arc is present with probability p, independently of the presence or absence of other arcs. We shall study the structure of the strongly connected components of digraphs drawn from $\boldsymbol{D}_{n,\,p}$, and shall give a very fast algorithm for constructing the transitive closure of such a digraph. All the results are asymptotic; i.e., they show that certain events hold almost certainly in the limit as n tends to infinity, with p varying in a prescribed way as a function of n.

The following is an informal sketch of the main results of the paper. Let D be a "typical" digraph drawn from $\boldsymbol{D}_{n,\,p}$. The structure of D depends on the quantity np, which gives the expected degree of a vertex. Let $X(r)$ denote the set of vertices reachable from vertex r, and let $Y(r)$ denote the set of vertices from which r is reachable (a vertex is defined to be reachable from itself). Let h be a fixed small positive constant. When $np < 1 - h$ each of the sets $X(r)$ is of size

† Research supported by NSF Grants DCR-8411954 and CCR-8612563, and by the International Computer Science Institute, Berkeley, California.

less than or equal to $B(h)\log n$, and the expected size of $X(1)$ is bounded above by a constant $C(h)$, where $B(h)$ and $C(h)$ are constants related to h. When np $> 1+h$ each of the sets $X(r)$ or $Y(r)$ is either of size less than $B(h)\log n$ or else contains a nonnegligible fraction of all the vertices; in the former case the set is called *small*, and in the latter case, *large*. The expected size of a small set is bounded above by $C(h)$. The $2n$ events "$X(r)$ is large" or "$Y(r)$is large" are nearly pairwise independent. Whenever $X(u)$ is large and $Y(v)$ is large there exists a path from u to v. The vertices r such that $X(r)$ is large and $Y(r)$ is large form a "giant strong component" of D containing a nonnegligible fraction of all the vertices. Each of the other strong components contains at most $B(h)\log n$ vertices. Nearly all the vertices outside the giant strong component lie in strong components of size 1. If *np* tends to infinity as n tends to infinity then the fraction of vertices lying in the giant strong component tends to 1.

When *np* remains equal to a constant $c > 1$ as n tends to infinity, more quantitative information can be given. The size of each large set $X(r)$ or $Y(r)$ is close to Θn, where Θ is the unique root of the equation $1-x-e^{-cx}=0$ in $[0,1]$. The probability that $X(1)$ is large is approximately Θ, and the size of the giant strong component is approximately $\Theta^2 n$.

Similar results hold when $np(n)=1+\varepsilon(n)$, where $\varepsilon(n)$ tends to 0 as n tends to infinity, but $\frac{\varepsilon(n)^3\, n}{\log^2 n} \underset{n\to\infty}{\longrightarrow} \infty$. In this case the size of a small set is bounded above by a constant times $\varepsilon(n)^{-2}\ \ln n$, and the giant strong component is of size approximately $2n\ \varepsilon(n)/(1+\varepsilon(n))^2$.

When $np < 1-h$ the number of pairs (u,v) in the transitive closure of D is $O(n)$, and the transitive closure can be computed in expected time $O(n)$ by conducting a breadth-first search from each vertex, provided that the arcs out of any vertex are accessible in a random order. When $np > 1+h$ the number of pairs (u,v) in the transitive closure is $\Omega(n^2)$, but the transitive closure can be represented in the compact form $(A\times B)\cup C$, where $A=\{u \mid X(u) \text{ is large}\}$, $B=\{v \mid Y(v) \text{ is large}\}$ and C is a set of $O(n)$ ordered pairs of vertices. An algorithm based on a combination of forward and backward breadth-first search computes this compact representation of the transitive closure of D in expected time $O(n)$. When *np* is extremely close to 1 the algorithm no longer runs in linear expected time. Instead, its expected execution time is $O(w(n)\ (n\log n)^{4/3})$ where $w(n)$ is an arbitrary nondecreasing unbounded function.

Throughout our investigation of the structure of random digraphs we emphasize proof techniques based on the analysis of search procedures for constructing the sets $X(r)$ and $Y(r)$. This algorithmically oriented approach is in contrast to the usual proof methods in the theory of random graphs, which are based on estimating the numbers of subgraphs with various properties. We believe that our approach leads to more straightforward proofs than the conventional methods do, and provides more insight.

Eddie Grove of the Computer Science Division at Berkeley has performed computations that provide insight into the rate at which the asymptotic behavior predicted by our theorems is approached. His program generated a set of random digraphs and, for each, computed the size of the giant strong component, the sizes of the sets A,B and C in the representation of the transitive closure as $(A\times B)\cup C$, and several other quantities of interest. The results of these computations are presented in Section 5.

Although there is an immense literature on the structure of random undirected graphs, random digraphs have been studied very little. An early result about the structure of the strong components is given in [Palásti, 1966], where it is shown that, as n tends to infinity with $np = \ln n + a$, the probability that D fails to be strongly connected tends to $\exp(-2\, e^{-a})$. Some futher results are given in [Łuczak, 1988]. An entirely different model of random digraphs is studied in [Fenner & Frieze, 1988]. The paper [Schnorr, 1985] presents an algorithm for constructing the transitive closure of a random digraph. For all n and p, Schnorr's algorithm runs in expected time $O(n^2)$.

2. THE GAP THEOREM

When $np < 1 - h$ it is likely that, in a random digraph D, every set $X(r)$ or $Y(r)$ is quite small. More interestingly, when $np > 1 + h$, a "gap phenomenon" occurs: it is likely that every set $X(r)$ or $Y(r)$ is either quite small or very large. We explore these phenomena in two ways: first, by introducing a rather general "conversion method" that allows certain results about random graphs to be converted directly to statements about random digraphs, and secondly, by a direct analysis of search methods for constructing the sets $X(r)$ and $Y(r)$.

2.1 *THE CONVERSION METHOD*

The theory of random graphs is largely concerned with probability spaces $\boldsymbol{G}_{n,p}$. The points in $\boldsymbol{G}_{n,\,p}$ are graphs (undirected, and without loops or multiple edges) with vertex set $\{1,2,...,n\}$. The probability assigned to a graph G with e edges is $p^e\,(1-p)^{\binom{n}{2}-e}$. Thus,in a graph G drawn from $\boldsymbol{G}_{n,\,p}$, each edge is present with probability p, independently of the presence or absence of other edges.

We shall show that certain known results about random graphs can be converted directly to results about random digraphs. As a bridging mechanism we introduce a family of probability spaces $\tilde{\boldsymbol{G}}_{n,\,p}$ of random digraphs. A digraph $\hat{G}$ can be drawn from $\tilde{\boldsymbol{G}}_{n,\,p}$ by the following experiment:

(i) draw a graph G from $\boldsymbol{G}_{n,\,p}$;

(ii) place the arcs $[u,v]$ and $[v,u]$ into $\tilde{G}$ if and only if G contains the edge $\{u,v\}$.

Lemma 1. Let G be drawn from $\boldsymbol{G}_{n,\,p}$, $\tilde{G}$ from $\tilde{\boldsymbol{G}}_{n,\,p}$ and D from $\boldsymbol{D}_{n,\,p}$. Then the following three random variables are identically distributed: the number of vertices in the connected component of G containing vertex 1, the number of vertices reachable from vertex 1 in $\tilde{G}$ and $X(1)$, the number of vertices reachable from vertex 1 in D.

Proof: The fact that the first two random variables are identically distributed follows immediately from the definition of $\tilde{\boldsymbol{G}}_{n,\,p}$ in terms of $\boldsymbol{G}_{n,\,p}$. To see that the last two random variables are identically distributed, note that the probability spaces $\tilde{\boldsymbol{G}}_{n,\,p}$ and $\boldsymbol{D}_{n,\,p}$ differ in only one respect: in a digraph G drawn from $\tilde{\boldsymbol{G}}_{n,\,p}$, arc $[u,v]$ is present if and only if arc $[v,u]$ is present, while, in a digraph D drawn from $\boldsymbol{D}_{n,\,p}$, the event that $[v,u]$ is present is independent of the event that $[u,v]$ is present. Thus, no experiment based on checking for the presence or absence of arcs can distinguish between the two probability spaces unless it checks both an arc and its reversal. But any standard sequential algorithm, such as breadth-first search or depth-first search, for building a search tree containing exactly the vertices reachable from vertex 1, checks for the presence of arc $[u,v]$ only if vertex u is in the search tree and v is not; thus it never checks both an arc and its reversal, and accordingly cannot distinguish $\tilde{\boldsymbol{G}}_{n,\,p}$ from $\boldsymbol{D}_{n,\,p}$. □

Lemma 1 allows us to convert results about the connected components of random graphs to results about reachability sets in digraphs. For example, the following results are either proved in [Bollobás, 1985] or can easily be extracted from results proved therein.

Let h be a (small) positive constant. Let $w(n)$ be a nondecreasing unbounded function. Let Θ be the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$. Let the size of a connected component be the number of vertices it contains.

Lemma 2. Let $p(n)$ be such that, for all n, $np(n) < 1 - h$. Then, with probability tending to one as n tends to infinity, a graph G drawn from $\boldsymbol{G}_{n,\,p}$ has all its connected components of size less than $3(ln\ n)h^{-2}$.

Lemma 3. Let $p(n)$ be such that, for all n, $np(n) > 1 + h$. Then, with probability tending to 1 as n tends to infinity, a graph G drawn from $\boldsymbol{G}_{n,\,p(n)}$ has exactly one connected component of size greater than $3(ln\ n)h^{-2}$.

Lemma 4. Let $p(n) = c/n$, where c is a constant greater than 1. Let G be drawn from $\boldsymbol{G}_{n,\,p(n)}$. Then, with probability tending to 1 as n tends to infinity, the size of the largest component of G lies between $\Theta n - w(n)\sqrt{n}$ and $\Theta n + w(n)\sqrt{n}$, where Θ is the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$.

Applying Lemma 1, we obtain the following results about the distribution of $X(1)$, the number of vertices reachable from vertex 1 in a digraph drawn from $\boldsymbol{D}_{n,\,p}$.

Corollary 1. Let h be a positive constant, $w(n)$ a nondecreasing unbounded function and Θ the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$. If $np(n) < 1 - h$ then, with probability tending to 1 as n tends to infinity, $|X(1)| < 3(ln\ n)\,h^{-2}$. If $p(n) = c/n$, where $c > 1$ then, with probability tending to 1 as n tends to infinity, $|X(1)|$ lies in the union of the two intervals $[0,\ 3(ln\ n)\,h^{-2}]$ and $[\Theta n - w(n)\sqrt{n},\ \Theta n + w(n)\sqrt{n}]$; moreover, the probability that $|X(1)|$ lies in the interval $[\Theta n - w(n)\sqrt{n},\ \Theta n + w(n)\sqrt{n}]$ tends to Θ as n tends to infinity.

2.2 *STRONGER FORMS OF THE GAP THEOREM*

Corollary 1 demonstrates an important gap phenomenon: when $np = c > 1 + h$ the number of vertices reachable from a given vertex is likely to be either very small (in the interval $[0,\ 3(ln\ n)\,h^{-2}]$) or very large (in the interval $[\Theta n - w(n)\sqrt{n},\ \Theta n + w(n)\sqrt{n}]$). The following theorem is a more quantitative version of Corollary 1; it gives upper bounds of the form n^{-a} on the probability that $|X(1)|$ does not lie in a prescribed union of two intervals. The proof is based on the analysis of a natural search process for constructing the set $X(1)$. A closely related process is studied in [Nagaev & Startsev, 1970].

Theorem 1. Let c be a constant greater than 1. Let a be a positive constant. Let B be a constant greater than $(a+1)\,c\,(c-1)^{-2}$. Let $w(n)$ be a nondecreasing unbounded function. Let Θ be the unique root in $[0,1]$ of the equation $1-x-e^{-cx}=0$. Let D be drawn from $\boldsymbol{D}_{n,\,c/n}$, and let $X(1)$ be the set of vertices reachable from vertex 1 in D. Then, for all sufficiently large n, $Pr[\,|X(1)|\notin[0,\,B\ln n\,]\cup[\Theta n-w(n)\sqrt{n},\,\Theta n+w(n)\sqrt{n}\,]<n^{-a}$

Proof: We consider a natural "fanning-out process" for constructing the set $X(1)$. The process constructs a sequence $<(A_0,B_0),\,(A_1,B_1),...,(A_t,B_t),...>$ where $A_i\subseteq B_i\subseteq\{1,2,...,n\}$ and $|\,A_i\,|=i$. The set B_i consists of the vertices that have been reached during the first i iterations of the process, and the set A_i consists of the vertices that have been scanned during the first i iterations. Here $A_0=\phi$, $B_0=\{1\}$ and the pair $(A_{i+1},\,B_{i+1})$ is constructed from $(A_i,\,B_i)$ by the following rule: $A_{i+1}=A_i\cup\{v\}$ and $B_{i+1}=B_i\cup succ(v)$ where v is a randomly chosen element of $B_i\backslash A_i$ and vertex w lies in the set $succ(v)$ if and only if the digraph D contains the edge (v,w). The process terminates when, for some t, $A_t=B_t$; i.e., termination occurs when every vertex that has been reached has also been scanned. It is clear that, if the process terminates after t iterations, then $A_t=B_t=X(1)$.

At each iteration of the fanning-out process, each vertex not already reached has probability p of being reached. This observation permits a simple description of the stochastic behavior of the fanning-out process. Let $BIN(\,n,\,p)$ be the binomial distribution with parameters n and p; i.e., the probability distribution of the number of heads in n independent coin tosses with probability of heads p. Then the conditional distribution of $|B_{i+1}\backslash B_i\,|$, given $B_0,\,B_1,...,B_i$, is $BIN(n-|B_i\,|,\,p)$ (of course, B_{i+1} is defined only if B_i is defined and $|B_i\,|>i$). It follows that the stochastic behavior of the fanning-out process can be described by a sequence of random variables $\{Z_i\}$, where $Z_0=1$ and $Z_{i+1}=Z_i+BIN(n-Z_i,\,p)$. The sequences $\{B_i\}$ and $\{Z_i\}$ have the same distribution, except that the former sequence terminates as soon as, for some t, $|B_t\,|=t$, while the latter sequence goes on forever. It follows that $|X(1)|$ has the same distribution as $\min\,\{t\mid Z_t=t\}$. This implies the inequality $Pr\,[\,|X(1)|=t\,]\le Pr\,[Z_t=t\,]$.

In view of the inequality $Pr\,[X(1)=t\,]\le Pr\,[Z_t=t\,]$, we study the probability distribution of Z_t in order to prove an upper bound on the probability that $|X(1)|=t$. It is easy to prove by induction that Z_t-1 has the probability distribution $BIN(n-1,\,1-(1-p)^t)$. Intuitively, this is because Z_t is intended to represent the number of vertices reached in the first t iterations of the fanning-out process, and the chance that a given vertex (other than vertex 1) gets reached during the first t iterations is $1-(1-p)^t$, which is the probability of at least one head in t independent tosses of a coin with probability of heads p.

We shall require the following two bounds on the tail of the binomial distribution [Raghavan, 1986]. Let the random variable X have the distribution $BIN(\,n,\,p)$. Then,

(2.1) For every positive real β, $Pr\,[X>\beta np\,]\le\left(\dfrac{e^{\beta-1}}{\beta^{\beta}}\right)^{np}$;

(2.2) For $\gamma\in[0,1]$, $Pr\,[X<(1-\gamma)\,np\,]<e^{-\gamma^2 np/2}$

We shall use these inequalities to get an upper bound on the probability that $Z_t = t$. It will be convenient to define the auxiliary sequence of random variables $\{W_i\}$ as follows: $W_0 = 1$ and $W_{i+1} = W_i + BIN(n - t, p)$. It is easily shown by induction that, for all $j \le t$, and all i, $Pr\,[Z_j \le i\,] \le Pr\,[W_j \le i\,]$. Hence, $Pr\,[X(1) = t\,] \le Pr\,[W_t \le t\,]$. But $W_t - 1$ clearly has the distribution $BIN(t(n-t), p)$, and it follows from inequality 2.2 that $Pr\,[W_t \le t\,] \le \exp\left(-((c-1)\,t - \frac{ct^2}{n})^2/ct\right)$. A brief calculation shows that, for any $B > \frac{(a+1)c}{(c-1)^2}$, and for all n sufficiently large, $Pr\,[W_t \le t\,] \le n^{-(a+1)}$ for all t in the interval $\left[B \ln n,\ \frac{c-1}{3}\, n \right]$.

It remains to consider the case where $t > \frac{c-1}{3}\, n$. We use the fact that $Z_t - 1$ has the distribution $BIN\left(n-1,\ 1 - \left(1 - \frac{c}{n} \right)^{xn} \right)$, where $x = t/n$. Using the inequalities

(i) for all real z, $\ln(1-z) < -z$ and

(ii) for $0 < z < .69$, $\ln(1-z) > -z - z^2$, we obtain

$$Pr\,[Z_t \le t\,] \le \exp\left(-n \frac{(1 - e^{-(cx + c^2 x/n)} - x)^2}{2(1 - e^{-cx})} \right)$$

and a brief calculation shows that, when $t = xn$, $x \le \Theta - w(n)\, n^{-1/2}$ and n is sufficiently large, $Pr\,[Z_t \le t\,] < n^{-(a+1)}$; here, Θ is the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$, and $w(n)$ is an arbitrary nondecreasing unbounded function.

A similar calculation based on the inequality (2.1) shows that, when $t = xn$, $x > \Theta + w(n)\, n^{-1/2}$ and n is sufficiently large, $Pr\,[Z_t \ge t\,] < n^{-(a+1)}$.

We have now shown that, for n sufficiently large, and for all t outside the intervals $[0, B \ln n\,]$ and $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}\,]$, $Pr\,[Z_t = t\,] < n^{-(a+1)}$. It follows that, when n is sufficiently large, the probability that $|Z(1)|$ lies outside the union of these two intervals is less than n^{-a}

□

We continue to consider digraphs drawn from $\boldsymbol{D}_{n,\, c/n}$, where $c > 1$. Let us say that $X(1)$, the set of vertices reachable from vertex 1, is *large* if $|X(1)|$ lies in the interval $[\Theta n - w(n)\sqrt{n}, \Theta n + w(n)\sqrt{n}\,]$, and *small* otherwise. Theorem 1 tells us that, if $X(1)$ is small, then, with overwhelming probability, its cardinality is bounded above by a small multiple of $c\,(c-1)^{-2} \log n$.

Theorem 2. Let D be drawn from $\boldsymbol{D}_{n,\, c/n}$ where $c > 1$. Then, as n tends to infinity,

(i) The probability that $X(1)$ is small tends to $1 - \Theta$, where Θ is the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$;

(ii) The expected size of $|X(1)|$, given that $X(1)$ is small, tends to $\frac{1}{1-c(1-\Theta)}$.

Proof: We shall analyze the early stages of the fanning-out process for constructing the set $X(1)$. Recall that a run of this process is described by a sequence $\{(A_i, B_i)\}$, where A_i is the set of vertices scanned at or before the ith iteration, and B_i is the set of vertices reached at or before the i^{th} iteration. Let A be a positive constant, and consider the sequence of iterations in which $|B_i| \le A \ln n$; this sequence of iterations will be called the *infancy* of the process. The number of vertices reached for the first time at iteration $i+1$ has the probability distribution $BIN(n-1-|B_i|,\ c/n)$. Since $|B_i| \le A \ln n$, this distribution is closely approximated, when n is very large, by the probability distribution $BIN(n-1,\ c/n)$. This suggests that the evolution of the fanning-out process during its infancy can be closely approximated by a branching process (cf. [Harris], [Athreya & Ney]) which starts with a single progenitor, and in which the number of children of each individual, independently of the behavior of all other individuals, has the distribution $BIN(n-1,\ c/n)$. We shall refer to this process as the *binomial branching process*. In this process, let us say that an individual is *mortal* if his total number of descendants is finite. Let q_n be the probability that the progenitor (or any individual) is mortal, and let s_n be the expected number of descendants of the progenitor (including the progenitor himself), given that the progenitor is mortal. To determine the behavior of q_n and s_n as n tends to infinity, note that $BIN(n-1,\ c/n)$ converges in distribution to the Poisson distribution with mean c. This suggests consideration of a *Poisson branching process* in which the number of children of any individual, independently of the behavior of all other individuals, has the Poisson distribution with mean c. Let q be the probability that the progenitor of this process is mortal, and let s be the expected number of descendants of the progenitor, given that the progenitor is mortal. In the Poisson branching process the conditional probability that an individual is mortal, given that he has k children, is q^k. Unconditioning, we find that q is the unique root in $[0,1]$ of the equation $q = \sum_{k=0}^{\infty} e^{-c} \frac{c^k}{k} q^k$, which becomes, after some simplification, $q = e^{-c(1-q)}$; hence $q = 1 - \Theta$. Further calculation using Bayes' Theorem shows that, when we condition on the event that the progenitor is mortal, the number of children of the progenitor has the Poisson distribution with mean cq. It is easily verified that $cq < 1$, and it follows that

$$s = \sum_{h=0}^{\infty} (cq)^k = \frac{1}{(1-cq)} = \frac{1}{1-c(1-\Theta)}.$$

From the fact that $BIN(n,\ c/n)$ converges in distribution to the Poisson distribution with mean c it follows easily that $q_n \underset{n\to\infty}{\longrightarrow} q$ and $s_n \underset{n\to\infty}{\longrightarrow} s$. By Markov's inequality, the probability that the number of descendants of the progenitor exceeds $A \ln n$, given that the process is mortal, is bounded above by $s_n / A \ln n$, which tends to zero. Thus, the probability that the number of descendants of the progenitor exceeds $A \ln n$ tends to $\Theta = 1-q$ as n tends to infinity.

It remains for us to show that the same result holds for the fanning-out process that determines $X(1)$. The only difference between the fanning-out process and the binomial branching process is that, in the branching process, no two parents have a child in common, so that the process can be represented as a tree, while, in the fanning-out process, a node may be reached along two different paths. However, the probability that such a "collision" will occur during the infancy of the fanning-out process, up to the point where $A \ln n$ nodes have been reached, is $O(\frac{\log^2 n}{n})$, and it follows that the probability that $|X(1)| < A \ln n$ tends to Θ, and the expected size of $X(1)$, given that it is less than $A \ln n$, tends to $\frac{1}{1-c(1-\Theta)}$. □

The preceding theorems concern the case where *np*, the expected degree of a vertex, is bounded away from 1. We shall also be interested in cases where *np* converges to 1 from above as *n* tends to infinity. Let $\varepsilon(n)$ be a positive real function that tends to zero as *n* tends to infinity, such that $\frac{\varepsilon(n)^3 n}{\log^2 n} \underset{n\to\infty}{\longrightarrow} \infty$. When *n* is understood we often write ε instead of $\varepsilon(n)$. Let $p(n) = \frac{1+\varepsilon(n)}{n}$. Then we obtain the following analogues of Theorems 1 and 2.

Theorem 1′. Let $w(n)$ be a nondecreasing unbounded function. Then, for all positive *a*, the following holds for all sufficiently large *n*:

$$Pr\left[x(1) \notin [0,(a+2)\varepsilon^{-2}\ln n\,] \cup \left[\frac{2\varepsilon n}{(1+\varepsilon)^2} - w(n)\sqrt{\frac{n\log n}{\varepsilon}},\ \frac{2\varepsilon n}{(1+\varepsilon)^2} + w(n)\sqrt{\frac{n\log n}{\varepsilon}}\right]\right] < n^{-a}$$

Theorem 2′. Let us say that $X(1)$ is *large* if $|X(1)|$ lies in the interval

$$\left[\frac{2\varepsilon n}{(1+\varepsilon)^2} - w(n)\sqrt{\frac{n\log n}{\varepsilon}},\ \frac{2\varepsilon n}{(1+\varepsilon)^2} + w(n)\sqrt{\frac{n\log n}{\varepsilon}}\right]$$

and *small* otherwise. then, as *n* tends to infinity,

(i) The probability that $X(1)$ is small is $1 - \frac{2\varepsilon}{(1+\varepsilon)^2} + o(\varepsilon^2)$

(ii) The expected size of $|X(1)|$, given that $X(1)$ is small, is $\frac{1+O(\varepsilon)}{\varepsilon}$.

The proofs of these theorems are quite similar to the proofs of Theorems 1 and 2. The details are omitted.

3. THE STRUCTURE OF THE TRANSITIVE CLOSURE

We continue to consider digraphs drawn from $\boldsymbol{D}_{n,\,c/n}$, where $c > 1$. Let *uRv* mean that vertex *v* is reachable from vertex *u*. Recall that $X(r)$ is the set of vertices reachable from vertex *r*, and $Y(r)$ is the set of vertices from which vertex *r* can be reached. In view of the gap theorem there is a constant *A* such that, with probability tending to 1, the cardinality of every set $X(r)$ or $Y(r)$ is either less than $A \ln n$ or else differs from Θn by at most $w(n)\sqrt{n}$. A set $X(r)$ or $Y(r)$ is said to be *small* in the former case and *large* in the latter case. By Theorem 2, the expected size of $X(r)$, given that $X(r)$ is small, is bounded above by a constant independent of n.

Theorem 3. Let D be drawn from $\mathbf{D}_{n,\ c/n}$, where $c > 1$. Then with probability tending to 1, the following statement holds for all u and v: if $X(u)$ is large and $Y(v)$ is large then uRv.

Proof: Since there are only n^2 choices for the ordered pair (u,v), it suffices to prove that, for any fixed pair (u,v) $Pr\,[X(u)$ is large $\cap\ Y(v)$ is large $\cap\ u \not R v] = o(n^{-2})$. This probability is bounded above by $Pr\,[X(u)$ is large $\cap\ Y(v)$ is large $\cap\ u \not R v]$. We shall prove that this conditional probability is $o(n^{-2})$. Consider the fanning-out process for constructing the set $X(u)$. Since we are given that $u \not R v$, there is no edge from a vertex in $X(u)$ to a vertex in $Y(v)$. Thus, whenever the process fans out from a vertex, the number of new vertices reached has a distribution stochastically smaller than $BIN(n - |Y(v)|,\ c/n)$. For any N, the probability that the process reaches as many as $N+1$ vertices is bounded above by the probability that at least N vertices are reached during the first N iterations of the fanning-out process; but the probability distribution of the number of vertices reached during N iterations is stochastically smaller than $BIN(N(n - |Y(v)|),\ c/n$. Taking $N = \Theta n - w(n)\sqrt{n}$, noting that $|Y(v)| \geq \Theta n - w(n)\sqrt{n}$ and $c(1 - \Theta) < 1$, and applying the bound (2.1), we find that $Pr\,[X(u)$ is large $|\ Y(v)$ is large and $u \not R v]$ is bounded above by a function that tends exponentially to zero, and thus is certainly $o(n^{-2})$. □

Let $A(u)$ be the event that $X(u)$ is large, and let $B(v)$ be the event that $Y(v)$ is large. We shall show that these $2n$ events are nearly pairwise independent. It follows from a general result given in [Harris] about the positive correlation of monotone properties that any two of these events are positively correlated. Thus, for example, for any two vertices u and v,

$$Pr\,[A(u)\ |\ \neg\, A(v)] \leq Pr\,[A(u)] \leq Pr\,[A(u)\ |\ A(v)] \qquad (2.3)$$

The following theorem shows that the positive correlation is not very strong.

Theorem 4. Let E and F be any two distinct events from $\{A(u)\} \cup \{B(v)\}$. Then

$$Pr\,[E\ |\ \neg F\,] \geq Pr\,[E\,] - O(log\ n/n) \qquad (2.4)$$

Proof: Consider, for example, the case where $E = A(u)$ and $F = A(v)$, $u \neq v$. We shall show that, for every small set X that does not contain u, $Pr\,[A(u)\ |\ X(v) = X\,] \geq Pr\,[A(u)] - O(\frac{\log n}{n})\ |\,X\,|$. Consider an experiment to determine, by a fanning-out process, whether $X(u)$ is large. This experiment terminates as soon as $A \ln n$ vertices are reached, and the experiment is unaffected by the information that $X(v) = X$ unless a vertex in X is reached during the process. But the probability that a random set $X(u)\ u$ of size at most $A \ln n$ intersects X is at most $\frac{A \ln n\ |X\,|}{n}$. The theorem now follows from two facts: given that $X(v)$ is small, the expected size of $X(v)$ is bounded above by a constant, and the probability that u lies in $X(v)$ is $O(1/n)$. The remaining cases are argued similarly.

A similar argument shows that any three or four of the above $2n$ events are nearly mutually independent. We now derive several corollaries pertaining to $\boldsymbol{D}_{n,\, c/n}$ where $c > 1$.

Corollary 2. Let E and F be any two distinct events from $\{A(u)\} \cup \{B(v)\}$. Then $Pr\,[E \mid F\,] \le Pr\,[E\,] + O\,(log\; n/n)$.

Proof: The result follows from Theorem 4 together with the inequality $Pr\,[E\,] = Pr\,[E \mid F\,]\, Pr\,[F\,] + Pr\,[E \mid \overline{F}\,]\, Pr\,[\overline{F}\,]$ and the fact that $Pr\,[E\,]$, $Pr\,[F\,]$ and $Pr\,[\overline{F}\,]$ all tend to positive constants as n tends to infinity.

Corollary 3. Let $LARGEOUT = \{u \mid X(u) \text{ is large}\}$, $LARGEIN = \{v \mid Y(v) \text{ is large}\}$ and $LARGE = \{u \mid X(u) \text{ is large and } Y(u) \text{ is large}\}$. Let $w(n)$ be a nondecreasing unbounded function. Then, with probability tending to 1, $|\, LARGEOUT - \Theta n \,| < w(n)\, \sqrt{n \log n}$, $|\, LARGEIN - \Theta n \,| < w(n)\, \sqrt{n \log n}$ and $|\, LARGE - \Theta^2 n \,| < w(n)\, \sqrt{n \log n}$.

Proof: The proofs of the three results are similar. As an example, consider the random variable $LARGEOUT$, which is equal to $\sum\limits_{u=1}^{n} A(u)$. Theorem 2 tells us that $E\,(A(1))$ tends to the limit Θ; the proof can be strengthened to show that $E\,[A(1)] = \Theta + O(n^{-1})$. It follows that $E\,[\, LARGEOUT\,] = \Theta n + O(1)$. Also, $E\,[\,(LARGEOUT)^2\,] = E\,[\,LARGEOUT\,] + n(n-1)\, E\,[A(u)\, A(v)]$, where u and v are any two distinct vertices. Applying Corollary 2,

$$E\,[\,A(u)\,A(v)\,] = E\,[\,A(u)\,] \cdot E\,[\,A(v) \mid A(u)\,] = E\,[\,A(u)\,]\,(E\,[\,A(v)\,] + O\,(\,\log n/n\,)\,].$$

It follows that the variance of $LARGEOUT$ is $O\,(n \log n)$, and the desired conclusion now follows from Chebyshev’s Inequality. □

Theorem 3 and Corollary 3 combine to yield the following fundamental Giant Strong Component Theorem for digraphs.

Theorem 5. Let $w(n)$ be a nondecreasing unbounded function. Let c be a constant greater than 1. There is a constant A such that, with probability tending to 1, a digraph drawn from $\mathbf{D}_{n,c/n}$ has exactly one strong component with more than $A \log n$ vertices, and the number of vertices in that strong component differs from $\Theta^2 n$ by at most $w(n)\, \sqrt{n \log n}$.

Proof: Any strong component containing a vertex not in $LARGE$ must be of size less than $A \log n$. By Theorem 5, it will be true with probability tending to 1 that, for every pair u,v of vertices in $LARGE$, uRv. In this case, the set $LARGE$ is a strong component. The result now follows from Corollary 3. □

Boris Pittel has pointed out that our proof methods can be adapted to to obtain a simple proof of Lemma 4, the classic Giant Component Theorem for random graphs. Since undirected graphs are not our main concern, we merely sketch this development. The main steps correspond to Theorem 1, Theorem 2, Corollary 3 and Theorem 5 of the present paper. We consider graphs drawn from $\mathbf{G}_{n,c/n}$, where $c > 1$. Let $X(u)$ denote the number of vertices in the connected component of vertex u. Applying our conversion principle (Lemma 1), Theorem 1 tells us that, with high probability, $|X(1)|$ is either *small* (i.e., less than $A \ln n$) or *large* (i.e., close to Θn), and Theorem 2 tells us that the probability that $X(1)$ is small tends to $1 - \Theta$. A proof similar to that of Corollary 3 than tells us that, with high probability, the number of vertices u with $X(u)$ large is close to Θ. This tells us that there are about Θn vertices in components of size about Θn, with all the other vertices lying in much smaller components, and it follows that $\{u \mid X(u) \text{ is large}\}$ must consist of a single component of size about Θn. This proof is somewhat simpler than earlier proofs of the same result (cf. [Bollobás, 1985]).

The following theorem states that nearly all the vertices of a digraph drawn from $\mathbf{D}_{n,c/n}$ can be expected to lie either in the giant strong component or in a strong component of size 1.

Theorem 6. Let c be a constant greater than 1. Let Θ be the unique root in $[0,1]$ of the equation $1 - x - e^{-cx} = 0$. In digraph D, call a vertex *exceptional* if it lies in a strong component that is of size greater than 1 and is not the unique largest strong component. Let the random variable s be the number of exceptional vertices in a digraph drawn from $\mathbf{D}_{n,c/n}$. Then the expected value of s is bounded above by a function of n that converges to the positive constant $\frac{2c^2(1-\Theta)^2}{1-c(1-\Theta)} - \frac{c^2(1-\Theta)^4}{1-c(1-\Theta)^2}$.

Proof: We may assume that the largest strong component is unique, that it consists of all those vertices u such that $X(u)$ and $Y(u)$ are both large, and that all other strong components are of size less than $A \log n$, where A is a constant; the contribution to the expected value of s of the rare digraphs that violate this condition is $o(1)$. Vertex v is exceptional if and only if it lies in a directed cycle C of length between 2 and $A \log n$ such that either (a) for all w in C, $X(w)$ is small or (b) for all w in C, $Y(w)$ is small. The expected number of vertices in cycles of each type is bounded above by

$$\sum_{t=2}^{A\log n} \frac{n(n-1)...(n-t+1)}{t} \left(\frac{c}{n}\right)^t \cdot t\, P_{n,t},$$

where $P_{n,t}$ is the probability that, for each vertex w in a given cycle of length t, $X(w)$ is small. Here $\frac{n(n-1)...(n-t+1)}{t}$ is the number of cyclic sequences of t distinct elements from $\{1,2,...,n\}$, $\left(\frac{c}{n}\right)^t$ is the probability that all the edges of such a cyclic sequence are present in the digraph D, and $P_{n,t}$ is the probability that, for all w in such a cyclic sequence, $X(w)$ is small. This latter probability can be shown to be $(1-\Theta)^t + t^2\, O(\log^2 n/n)$ by a branching-process argument similar to the proof of Theorem 2; the only modification is that the zeroth generation of the branching process is taken to have t elements (the elements of the cycle) rather than one. It follows that

$$\sum_{t=2}^{A\log n} \frac{n(n-1)...(n-t+1)}{t} \left[\frac{c}{n}\right]^t t\, P_{n,t} = \frac{c^2\,(1-\Theta)^2}{1-c\,(1-\Theta)} + o\,(1).$$

We must also consider directed cycles c of length between 2 and $A \log n$ such that, for each vertex w in c, both $X(w)$ and $Y(w)$ are small. Similarly, the expected number of vertices in such cycles is $\frac{c^2(1-\Theta)^4}{1-c\,(1-\Theta)^2}$. the result now follows by the Principle of Inclusion and Exclusion. □

The foregoing results provide information about the probable structure of the transitive closure of a digraph drawn from $\mathbf{D}_{n,\,p}(n)$ when $np\,(n)$ is equal to a constant c greater than 1. Corresponding results can be proven by similar methods when $p\,(n) = \frac{1+\varepsilon(n)}{n}$, where $\varepsilon(n)$ tends to zero and $\frac{\varepsilon(n)^3\, n}{\log^2 n} \xrightarrow[n\to\infty]{} \infty$. These results are stated below without proof. In the statements we often abbreviate $\varepsilon(n)$ by ε. Recall that, in this context, $X\,(u)$ is said to be *large* if it lies in the interval

$$\left[\frac{2\varepsilon n}{(1+\varepsilon)^2} - w\,(n)\sqrt{\frac{n\log n}{\varepsilon}},\ \frac{2\varepsilon n}{(1+\varepsilon)^2} + w\,(n)\sqrt{\frac{n\log n}{\varepsilon}}\right],$$

and *small* otherwise. Also, $A\,(u)$ denotes the event that $X\,(u)$ is large, and $B\,(v)$ denotes the event that $Y\,(v)$ is large.

Theorem 3′. Let D be drawn from $\mathbf{D}_{n,\,p}(n)$, where $p\,(n) = \frac{1+\varepsilon(n)}{n}$, where $\varepsilon(n)$ tends to zero and $\frac{\varepsilon(n)^3\, n}{\log^2 n} \xrightarrow[n\to\infty]{} \infty$. Then, with probability tending to 1, the following statement holds for all u and v: if $X\,(u)$ is large and $Y\,(v)$ is large then uRv.

Theorem 4′. Let E and F be any two distinct events from $\{A\,(u)\} \cup \{B\,(v)\}$. Then $Pr\,[E \mid \neg F\,] \ge Pr\,[E\,] - O\,(\varepsilon^{-1}/n)$.

Corollary 2′. Let E and F be any two distinct events from $\{A\,(u)\} \cup \{B\,(v)\}$. Then $Pr\,[E \mid F\,] \le Pr\,[E\,] + O\,(\varepsilon^{-2}/n)$.

Corollary 3′. Let $LARGEOUT = \{u \mid X(u) \text{ is large}\}$, $LARGEIN = \{v \mid Y(v) \text{ is large}\}$, and $LARGE = \{u \mid X(u) \text{ is large and } Y(u) \text{ is large}\}$. Let $w(n)$ be a nondecreasing unbounded function. Then, with probability tending to 1, $\left| LARGEOUT - \frac{2\varepsilon n}{(1+\varepsilon)^2} \right| \le w(n)\, n^{1/2}\, \varepsilon^{-1/2}$, $\left| LARGEIN - \frac{2\varepsilon n}{(1+\varepsilon)^2} \right| \le w(n)\, n^{1/2}\, \varepsilon^{-1/2}$, and $\left| LARGE - \frac{4\varepsilon^2 n^2}{(1+\varepsilon)^4} \right| \le w(n)\, n^{1/2}\, \varepsilon^{-1/2}$.

Theorem 5′. Let $w(n)$ be a nondecreasing unbounded function. Let D be drawn from $\mathbf{D}_{n,\,p}(n)$, where $p(n) = \frac{1+\varepsilon(n)}{n}$, where $\varepsilon(n)$ tends to zero and $\frac{\varepsilon(n)^3\, n}{\log^2 n} \xrightarrow[n \to \infty]{} \infty$. There is a constant A' such that, with probability tending to 1, D has exactly one strong component with more than $A'\, \varepsilon^{-2} \ln n$ vertices, and the number of vertices in that strong component differs from $4\varepsilon^2 n^2 / (1+\varepsilon)^4$ by at most $w(n)\, n^{1/2}\, \varepsilon^{-1/2}$.

Theorem 6′ Under the same assumptions as in Theorem 5′, let s be the number of vertices that lie in a strong component that is of size greater than 1 but is not the unique largest strong component. Then the expected value of s is $O(\varepsilon^{-1})$.

4. A TRANSITIVE CLOSURE ALGORITHM

In this section we present an algorithm for constructing the transitive closure of a digraph. The algorithm runs in expected time $O(n)$ on digraphs drawn from $\mathbf{D}_{n,\,p}$,provided that $|\, np - 1 \,| \ge h$, where h is an arbitrary small positive constant. If no restriction is placed on p, then the algorithm is guaranteed to run in expected time $O(w(n)\,(n \log n)^{4/3})$, where $w(n)$ is an arbitrary nondecreasing unbounded function.

The execution time of our algorithm is clearly bounded above by the time required to write down the output. Since the expected number of pairs in the transitive closure is $\Omega(n^2)$ when np is greater than $1+h$, we adopt a special "Cartesian product representation" for the transitive closure: this representation is of the form $(A \times B) \cup C$, where A and B are sets of vertices and C is a set of ordered pairs of vertices.

Let us check that this representation is sufficiently compact. Suppose first that $np = 1 + \varepsilon(n)$, where $\varepsilon(n) \ge w(n)\,(log\, n)^{2/3}\, n^{-1/3}$. By Theorem 3, we obtain a correct representation of the transitive closure, with high probability, by taking $A = \{u \mid X(u) \text{ is large}\}$, $B = \{v \mid Y(v) \text{ is large}\}$, and $C = \{(u,v) \mid (X(u) \text{ is small and } v\, \varepsilon\, X(u)) \text{ or } (Y(v) \text{ is small and } u\, \varepsilon\, Y(v))\}$ By Theorem 2′, the expected size of the set C is $O(\frac{n}{\varepsilon}) = O(w(n)\, n^{4/3}\, (\log n)^{2/3})$, and hence the expected length of the output is also $O(w(n)\, n^{4/3}\, (\log n)^{2/3})$. On the other hand, suppose that $np \le 1 + \varepsilon^*(n)$, where $\varepsilon^*(n) = w(n)^{1/2}\, (\log n)^{2/3}\, n^{-1/3}$. Then the expected size of the transitive closure is no greater than the expected size of the transitive closure when $np = 1 + \varepsilon^*(n)$; and in that case, Theorems 1′ and 2′ tell us that the expected size of the transitive closure is asymptotic to $\frac{4\varepsilon^*(n)^2}{(1+\varepsilon^*(n))^4}\, n^2 + \frac{n}{\varepsilon^*(n)}$, which is $O(w(n)\,(n \log n)^{4/3}$.

Our algorithm will involve two primitive operations, *NEXTSUCC*(u) and *NEXTPRED*(u), where u is a vertex. Each execution of *NEXTSUCC*(u) will return a vertex v drawn at random from the set of vertices that are reachable from u by a single edge and have not been returned in previous executions of *NEXTSUCC*(u); if no such vertex exists then *NEXTSUCC*(u) will return the special symbol $*$. Similarly, successive executions of *NEXTPRED*(u) will return random samples without replacement from the set of vertices from which u is reachable by a single edge. We assume that each execution of one of these primitive operations takes unit time.

4.1. *A METHOD OF COMPUTING THE SET OF VERTICES REACHABLE FROM A GIVEN VERTEX*

In preparation for giving the transitive closure algorithm, we present a procedure for computing $X(r)$, the set of vertices reachable from a given vertex r. This procedure finds elements of $X(r)$ one at a time. For all n, p and m, on digraphs drawn from $\boldsymbol{D}_{n,\,p}$, the expected time for the procedure to find $\min(m,\ |X(r)|)$ elements of $X(r)$ is bounded above by αm, where α is a constant independent of n, p and m.

It is instructive to understand why the most obvious fanning-out method does not run fast enough. The following procedure realizes a version of this method.

FORWARD BREADTH-FIRST SEARCH

The input is a digraph D and a root vertex r. The set X contains the vertices reached so far, and the first-in first-out queue Q denotes the set of vertices that have been reached but not yet completely scanned. When the procedure terminates, X is equal to $X(r)$.

```
$X \leftarrow \{r\}$; $Q \leftarrow$ empty queue;
insert $r$ into $Q$;
while $Q$ is not empty do
    $u \leftarrow$ first element of $Q$
    $v \leftarrow NEXTSUCC(u)$;
    if $v = *$ then delete first element of $Q$;
    if $v \notin X \cup \{*\}$
    then insert $v$ into $Q$; $X \leftarrow X \cup \{v\}$
output $X$
```

Consider the execution of the algorithm while $|X| = i$. Every time the operation *NEXTSUCC*(u) is executed, all vertices not previously found to be directly reachable from u are equally likely to be returned. Thus, given that *NEXTSUCC*(u) does not return $*$, the probability that it returns a vertex not previously reached is at least $(n-i)/(n-1)$. Hence, the expected number of vertices drawn while $|X| = i$ is at most $(n-1)/(n-i)$. Thus, the expected number of accesses to *NEXTSUCC* lists, up to the point where m vertices have been reached, is at most $\left[\sum_{i=0}^{m-1} (n-1)/(n-i)\right] + m$, where the second term is an upper bound on the number of accesses that return $*$. This expectation is only reduced by the possibility that $|X(r)|$ may be less than m, since, in that case, the computation terminates before m vertices have been reached.

The summation $\left[\sum_{i=0}^{m-1} (n-1)/(n-i) \right] + m$, is bounded above by a constant times m only when m is bounded away from n; when $m = n$, the summation is $\Theta(n \log n)$. Thus, the difficulty with FORWARD BREADTH-FIRST SEARCH occurs when m is close to n. To get around this difficulty we introduce an alternate algorithm that searches both forward and backward. In its first phase it conducts a forward breadth-first search from r until either the search terminates or the number of elements in the queue Q becomes greater than $.1n$. In the latter case, from each vertex s that did not enter the set of vertices X reached from r during the forward breadth-first search, the procedure conducts a backward breadth-first search, using the *NEXTPRED* operation, to determine the vertices from which s is reachable. As soon as s is found to be reachable from a vertex in X the backward search from s terminates, and s is inserted into X. The procedure terminates with X equal to $X(r)$.

HYBRID SEARCH

```
$X \leftarrow \{r\}$; $Q \leftarrow$ empty queue; insert $r$ into $Q$;
while $Q$ is not empty and $|Q| \le .1n$ do
    $u \leftarrow$ first element of $Q$;
    $v \leftarrow NEXTSUCC(u)$;
    if $v = *$ then delete first element of $Q$;
    if $v \notin X \cup \{*\}$
    then insert $v$ into $Q$; $X \leftarrow X \cup \{v\}$
if $|Q| > .1n$ then
    for all $s \notin X$ do
        $X' \leftarrow \{s\}$; $Q' \leftarrow$ empty queue; insert $s$ into $Q'$;
        while $Q'$ is not empty and $X \cap X' = \phi$ do
            $w \leftarrow$ first element of $Q'$;
            $x \leftarrow NEXTPRED(w)$;
            if $x = *$ then delete first element of $Q'$;
            if $x \notin X' \cup \{*\}$
            then insert $x$ into $Q'$; $X' \leftarrow X' \cup \{x\}$
        if $X \cap X' \neq \phi$ then $X \leftarrow X \cup \{s\}$
output $X$
```

Theorem 6. For all n, p and m, the expected execution time required by HYBRID SEARCH to find $\min(m, |X(r)|)$ elements of $X(r)$ in a digraph drawn from $\boldsymbol{D}_{n,p}$, is bounded above by αm, where α is an absolute constant.

Proof: At the point where backward searches are begun from the vertices not in X, each vertex u in Q except possibly the one at the head of Q is "virgin"; i.e., no accesses have been made to $NEXTSUCC(u)$, and thus no edge (u,x) directed out of u has been excluded from being present in the digraph. Now consider the backward search from a vertex s; each time a list $NEXTPRED(x)$

is

accessed, each vertex u in Q (except possibly the one at the head of Q) is eligible to be drawn (i.e., there is no conditional information that the edge (u,x) does not exist), and thus,at each step, the probability that the next vertex drawn lies in Q is at least $\frac{|Q|-1}{n}$, which is at least .1. It follows that the expected execution time of each backward search is $O(1)$, and hence the total expected execution time of the backward searches is $O(n)$. Since the backward searches occur only when $|X(r)| > .1n$, the expected execution time of the backward searches is bounded above by a constant times $|X(r)|$.

It remains to be shown that the expected execution time of the forward search, up to the point where $\min(m, |X(r)|)$ elements are reached or the number of vertices in Q exceeds $.1n$, is bounded above by αm. We break the analysis into three cases.

CASE 1: $m < .95n$. It suffices to note that the forward search conducted by the algorithm is a truncation of FORWARD BREADTH-FIRST SEARCH, and that, when $m < .95n$, FORWARD BREADTH-FIRST SEARCH finds $\min(m, |X(r)|)$ elements of $X(1)$ in expected time bounded by a uniform constant times m.

CASE 2: $m > .95n$ and $np < 3$. Let z be the unique root of $1 - x - e^{-3x} = 0$ in $[0,1]$. Then $z < .95$, and thus, except with exponentially small probability, the inequality $|X(r)| < .95n$ will hold. The contribution of the exponentially rare cases in which $|X(r)| > .95n$ is negligible, since the worst-case running time of FORWARD BREADTH-FIRST SEARCH is $O(n^2)$. Thus, this case reverts to CASE 1.

CASE 3: $m > .95n$ and $np > 3$. Consider the execution of FORWARD BREADTH-FIRST SEARCH up to the point where Q becomes empty or $.4n$ executions of the *NEXTSUCC* operation have returned vertices (rather than $*$). Straightforward analysis shows that each of the following two events has exponentially small probability:

(i) more than $.2n$ vertices get scanned during this period;

(ii) Q does not become empty during this period, so that $.4n$ executions of *NEXTSUCC* return vertices, but the number of distinct vertices reached is less than $.3n$.

Thus, except in exponentially rare cases, Q will either become empty or attain size $.1n$ by the time $.6n$ executions of *NEXTSUCC* have occurred (at most $.4n$ of these will return vertices, and at most $.2n$ will return $*$). The exponentially rare exceptional cases make a negligible contribution to the expected time for the forward search, since the worst-case running time of FORWARD BREADTH-FIRST SEARCH is $O(n^2)$. It follows that the expected execution time of the forward search is $O(m)$. □

4.2. THE MAIN ALGORITHM

We are now ready to present our transitive closure algorithm. The algorithm uses HYBRID SEARCH as a subroutine for computing $\min(m, \mid X(r) \mid)$ elements of $X(r)$, where m is a specified integer and r is a specified vertex; it also requires a "mirror image" procedure called REVERSE HYBRID SEARCH that is used to compute $\min(m, \mid Y(r) \mid)$ elements of $Y(r)$, the set of vertices from which vertex r is reachable. The execution of REVERSE HYBRID SEARCH on digraph D can be viewed as the execution of HYBRID SEARCH on the digraph obtained by reversing all the edges of D.

It is convenient to present our algorithm as a combination of three separate algorithms having different domains of applicability. Let $\varepsilon^*(n) = w(n)^{1/2} (\log n)^{2/3} n^{-1/3}$ and let h be a small positive constant. The first algorithm is applicable in the *low density case,* where $np(n) \le 1 + \varepsilon^*(n)$; the second is applicable in the *high density case,* where $np(n) \ge 1 + h$; and the third is applicable in the *intermediate density case* where $1 + \varepsilon^*(n) \le np(n) < 1 + h$. We shall refer to these three algorithms as the *low density, high density* and *intermediate density* algorithms. It is not difficult to dovetail the three algorithms into a single algorithm that runs within the claimed bounds on expected time for all values of n and p, and does not require the user to specify p.

The low density algorithm is particularly simple. It computes the transitive closure explicitly, using HYBRID SEARCH to determine each set $X(u)$. When $np < 1 - h$ the expected size of each set $X(u)$ is easily seen to be less than $1/(1 - np)$, which is bounded above by $1/h$. Thus the expected time for each search is $O(1)$, and the expected time for the entire algorithm is $O(n)$. When $1 - h < np < 1 + \varepsilon^*(n)$ the expected size of each set $X(u)$ is bounded above by the expected size of $X(u)$ when $np = 1 + \varepsilon^*(n)$; this latter expectation is equal to $Pr[X(u)$ is large$] \cdot E[X(u) \mid X(u)$ is large$] + Pr[X(u)$ is small$] \cdot E[X(u) \mid X(u)$ is small$]$ which, by Theorems 1′ and 2′, is asymptotic to

$$\frac{2\varepsilon^*}{(1+\varepsilon^*)^2} (2\varepsilon^* n) + \frac{1 - 2\varepsilon^*}{(1+\varepsilon^*)^2} \left(\frac{n}{\varepsilon^*}\right) = O(w(n) (n \log n)^{4/3})$$

It follows that, in the range $1 - h \le np(n) \le 1 + \varepsilon^*(n)$, the expected execution time of the algorithm for the low density case is $O(w(n) (n \log n)^{4/3})$.

In the case where $np > 1 + h$ the high density algorithm identifies a vertex r such that $X(r)$ and $Y(r)$ are both large, and then computes the transitive closure in the form $(Y(r) \times X(r)) \cup C(r)$, where $C(r) = \{ (u,v) \mid (u \notin Y(r)) \wedge (v \in X(u)) \cup ((v \notin X(r)) \wedge (u \in Y(v))$. This representation is useful for the following reason: if $(X(r)$ and $Y(r)$ are represented by n-bit arrays and $C(r)$ is represented by a lexicographically ordered list of ordered pairs then, using binary search, one can determine in time $O(\log n)$ whether any given ordered pair (u,v) lies in the transitive closure. The algorithm for this case is as follows.

HIGH DENSITY ALGORITHM

```
/$np > 1 + h$/
$m \leftarrow \lceil 4\, h^{-2} \ln n \rceil$; $r \leftarrow n$
for $i = 1$ to $n$ do
    find $\min(m, \mid X(i) \mid)$ elements of $X(i)$;
    if $\min(m, \mid X(i) \mid) = m$
    then find $\min(m, \mid Y(i) \mid$ elements of $Y(i)$;
        if $\min(m, \mid Y(i) \mid) = m$ then $r \leftarrow i$; go to EXIT

EXIT:   compute $X(r)$ and $Y(r)$;
        for $v \notin X(r)$ compute $Y(v)$;
        for $u \notin Y(r)$ compute $X(u)$
        output $(Y(r) \times X(r)) \cup C(r)$ where
          $C(r) = \{(u,v) \mid (u \notin Y(r)) \wedge (v \in X(u) \cup ((v \notin X(r)) \wedge (u \in Y(v)).$
```

Theorem 7. Let h be a positive constant. Let $g(n)$ be the maximum, over all p such that $np > 1 + h$, of the expected execution time of the high density algorithm on digraphs drawn from $\boldsymbol{D}_{n,\,p}$. Then $g(n) = O(n)$.

Proof: We shall show that each of the following is $O(n)$:

(i) the expected time spent in the **for** loop;

(ii) the expected time to compute $X(r)$ and $Y(r)$;

(iii) the expected time to compute $C(r)$.

The **for** loop terminates as soon as an index i is found such that $\mid X(i) \mid \geq m$ and $\mid Y(i) \mid \geq m$. Since $np > 1 + h$, the probability that the loop terminates in its first iteration is greater than or equal to the probability of termination in the first iteration when $np = 1 + h$; by Corollary 3, that latter probability is bounded below by a constant b. Each unsuccessful iteration of the **for** loop examines $2m$ vertices, and thus conditions the behavior of later iterations, and reduces the probability of termination. However, this reduction is not greater than the reduction that would occur if all vertices reached in unsuccessful iterations were deleted from the digraph. The probability that the $(i+1)^{th}$ iteration is affected by such deletions is $O(m^2\ i/n)$, and thus the probability of termination at iteration $i + 1$, given that iteration did not occur during the first i iterations, is greater than $a - O(m^2\ i/n)$. It follows that the expected number of iterations is $O(1)$; and, since the expected time per iteration is $O(m)$, the expected time spent in the **for** loop is $\Theta(n)$. By Theorem 6, the expected time to compute $X(r)$ and $Y(r)$ using HYBRID SEARCH is $O(n)$.

Finally, by Theorem 3, we may assume that v does not lie in $X(r)$ if and only if $Y(v)$ is small, and u does not lie in $Y(r)$ if and only if $X(u)$ is small. By Theorem 4, the expected size of a small set is $O(1)$, and thus the expected time for HYBRID SEARCH to compute each small set $Y(v)$ or $X(u)$ is $O(1)$; it follows that the expected time to compute $C(r)$ is $O(n)$. □

The intermediate density algorithm is identical with the high density algorithm, except that the variable m, defining the boundary between large and small reachability sets, is set to $\lceil 4(\varepsilon^{*}(n))^{-2} \ln n \rceil$ instead of $\lceil 4 h^{-2} \ln n \rceil$. By a proof quite similar to the proof of Theorem 7, we obtain the following theorem.

Theorem 7′. Let $k(n)$ be the maximum, over all p such that $1 + \varepsilon^{*}(n) \le np \le 1 + h$, of the expected execution time of the intermediate density algorithm on digraphs drawn from $\mathbf{D}_{n,\,p}$. Then $k(n) = O(w(n)\,(n \log n)^{4/3})$. □

5. COMPUTATIONAL RESULTS

In this section we describe computational results obtained by Eddie Grove of the Computer Science Division at Berkeley. His program generated random digraphs for various values of n and p, and computed the following quantities: s, the number of strong components, g, the size of the largest strong component, and a representation of the transitive closure in the form $(A \times B) \cup C$, where A is the set of vertices from which the largest strong component is reachable, and B is the set of vertices reachable from the largest strong component.

In Grove's first experiment n was varied while holding np fixed at various values $c > 1$. Our theoretical asymptotic results indicate that, when n is large and $c > 1$, the sets A and B should each be of cardinality close to Θn, and g, the size of the largest strong component, should be close to $\Theta^2 n$, where Θ is the unique root in $[0,1]$ of $1 - x - e^{-cx} = 0$. To facilitate comparison with the asymptotic results, we give the values of Θ and Θ^2 for each $c > 1$, and tabulate s and g, as well as the following normalized quantities: g/n, $|A|/n$ and $|B|/n$.

In Grove's second experiment n was varied while holding np equal to $1 + n^{-1/4}$. One hundred trials were carried out for each value of n. In this case our theoretical asymptotic results indicate that, when n is large, the sets A and B should each be of cardinality close to $2\,n^{3/4} / (1 + n^{-1/4})^2$, and s, the size of the largest strong component, should be close to $4\,n^{1/2} / (1 + n^{-1/4})^4$. To facilitate comparison with the asymptotic results, we tabulate, for each n, the average value of s, g, $|A|$, and $|B|$ over the one hundred trials, and the quantities $2\,n^{3/4} / (1 + n^{-1/4})^2$ and $4\,n^{1/2} / (1 + n^{-1/4})^4$.

The computational results agree well with the asymptotic theory; as expected, the convergence to the asymptotic behavior is slowest when c is near 1. The results also confirm Theorem 6, which predicts that nearly every vertex will lie either in the largest strong component or in a strong component of size 1.

c = 2	**Θ = .799**		**Θ² = .638**		
n	**s**	**g**	**g/n**	**|A|/n**	**|B|/n**
8	6	2	.250	.375	.750
16	7	9	.562	.875	.688
32	18	14	.438	.562	.750
64	30	35	.547	.781	.672
128	70	59	.461	.750	.625

256	72	185	.723	.855	.848
512	220	293	.572	.721	.795
1024	380	645	.630	.803	.785
2048	743	1306	.638	.798	.806
4096	1429	2668	.651	.800	.809
8192	2953	5240	.640	.793	.806
16384	5819	10566	.645	.799	.805
32768	12025	20742	.633	.799	.793
65536	24327	41210	.629	.793	.794

c = 1.8	Θ = .732		Θ^2 = .536		
n	**s**	**g**	**g/n**	**|A|/n**	**|B|/n**
8	8	1	.125	.125	.750
16	5	12	.750	.750	1.000
32	8	25	.781	.938	.844
64	30	35	.547	.797	.703
128	60	69	.539	.758	.703
256	129	127	.496	.742	.656
512	211	302	.590	.768	.762
1024	484	541	.528	.725	.730
2048	962	1087	.531	.732	.711
4096	2007	2090	.510	.715	.718
8192	3709	4483	.547	.743	.737
16384	7683	8702	.531	.730	.728
32768	15330	17439	.534	.729	.730
65536	30495	35042	.535	.732	.730

c = 1.6	Θ = .642		Θ^2 = .412		
n	**s**	**g**	**g/n**	**|A|/n**	**|B|/n**
8	7	2	.125	.129	.125
16	7	10	.438	.562	.688
32	21	12	.469	.688	.719
64	50	15	.328	.453	.531
128	121	5	.547	.766	.742
256	158	98	.297	.496	.652
512	339	173	.504	.662	.734
1024	711	314	.448	.699	.636
2048	1652	397	.394	.660	.604
4096	2926	1171	.409	.627	.649
8192	6273	1910	.425	.648	.649
16384	12090	4293	.428	.645	.664
32768	24555	8214	.414	.639	.645
65536	48486	17038	.412	.645	.638

c = 1.4	**Θ = .512**		**Θ^2 = .262**		
n	**s**	**g**	**g/n**	**|A|/n**	**|B|/n**
8	7	2	.250	.750	.375
16	7	10	.625	.688	.938
32	21	12	.375	.594	.625
64	50	15	.234	.344	.391
128	121	5	.039	.539	.078
256	158	98	.383	.547	.684
512	339	173	.338	.566	.604
1024	711	314	.307	.566	.529
2048	1652	397	.194	.457	.440
4096	2926	1171	.286	.527	.537
8192	6273	1910	.233	.475	.501
16384	12090	4293	.262	.512	.515
32768	24555	8214	.251	.504	.502
65536	48486	17038	.260	.512	.510

c = 1.2	**Θ = .311**		**Θ^2 = .097**		
n	**s**	**g**	**g/n**	**|A|/n**	**|B|/n**
8	7	2	.250	.250	.750
16	11	3	.188	.250	.875
32	32	1	.031	.031	.031
64	58	7	.109	.375	.188
128	91	25	.195	.547	.414
256	237	12	.047	.211	.316
512	506	7	.014	.062	.281
1024	947	75	.073	.318	.239
2048	1882	164	.080	.359	.250
4096	3600	485	.118	.345	.330
8192	7580	587	.072	.266	.280
16384	14906	1474	.090	.294	.312
32768	29567	3201	.098	.323	.310
65536	59002	6529	.100	.307	.326

Table 1. Results of First Experiment

	Average Values					
n	**s**	**g**	**|A|**	**|B|**	**$2n^{3/4}/(1+n^{-1/4})^2$**	**$4n^{1/2}/(1+n^{-1/4})^4$**
8	5	3	5	4	3.74	1.75
16	10	5	10	8	7.10	3.15
32	25	7	15	13	13.3	5.57
64	50	13	27	26	24.7	9.54
128	107	18	49	43	45.3	16.0
256	220	33	88	85	81.9	26.1
512	460	46	149	139	147	42.0
1024	950	67	254	239	262	67.6
2048	1923	115	470	459	463	104
4096	3916	167	807	802	807	160
8192	7914	260	1416	1453	1409	246
16384	15981	383	2450	2409	2441	360
32768	32186	557	4171	4234	4194	524
65536	64632	881	7479	7294	7209	786

Table 2. Results of Second Experiment

6. ACKNOWLEDGEMENTS

The author thanks Eddie Grove, who performed the experimental computations reported in Section 5, Donald Knuth and Boris Pittel, who listened carefully to an early version of the results and provided perceptive suggestions for further research, and Robert Tarjan, who provided useful comments in the early stages of the work.

7. REFERENCES

K. B. Athreya, P. E. Ney, *Branching Processes,* Springer-Verlag (1972).

B. Bollóbas, *Random Graphs,* Academic Press (1985).

T. I. Fenner, A. M. Frieze, "On the connectivity of random m-orientable graphs and digraphs," *Combinatorica,* vol. 2, 347-359 (1988).

T. E. Harris, *The Theory of Branching Processes,* Springer (1963).

T. Łuczak, "The phase transition in the evolution of random graphs," manuscript (1988).

A. V. Nagaev, A. N. Startsev, "The asymptotic analysis of a stochastic model of an epidemic," *Theory of Probability and its Applications,* vol. XV, No. 1, pp. 98-107 (1970).

I. Palásti, "On the strong connectedness of directed random graphs," *Studia Sci. Math. Hungar.,* vol. 1, pp. 205-214 (1966).

P. Raghavan, "Probabilistic Construction of deterministic algorithms: approximating packing integer programs," *Proc. 27th Annual IEEE Symp. on Foundations of Computer Science,* pp. 10-18 (1986).

C. P. Schnorr, "An algorithm for transitive closure with linear expected time," *SIAM J. Comput.,* vol. 7, 127-133 (1978).