# Giant Components for Two Expanding Graph Processes

Luc Devroye, Colin McDiarmid, Bruce Reed

**ABSTRACT:** *We discuss the emergence of giant components in two random graph models (one directed, one undirected). Our study of these models was motivated by an interest in finding a random model of the Internet.*

## 1 Introduction

The hyperlinks between the pages of the internet yield a directed graph whose vertices are the web pages and whose arcs correspond to the hyperlinks themselves.

This directed graph and the undirected graph underlying it have been intensely studied (see Adamic and Huberman 1999, Broder et al, 2000, Kleinberg et al. 1999) as an understanding of its structure could be useful in designing searching engines or identifying communities on the web. Researchers are also attempting to build random models of the web (see Barabási Albert and Jeong 1999, Cooper and Frieze 2001, Kumar et al. 1999).

As pointed out in (Kumar et al. 1999), standard random graph models do not accurately represent the web for two reasons. The first is that the web has more vertices of high degree than an average graph. The second is that the web expands as pages get added over time, and a page is more likely to link to those which were present when it was added.

Indeed, this expansion is to some extent responsible for the existence of high degree vertices, as old pages tend to have high degree. However, the function a page serves is also important in determining its degree. For example, the home page for Google has very high degree.

Researchers (see Aiello Chung and Lu 2000, Strogatz and Watts 1999) have applied the techniques of (Molloy and Reed 1995), to study the connectivity properties of graphs whose degree sequence is similar to that of the undirected graph underlying the web. However, less attention has been devoted to developing models which reflect the time dependency inherent in the internet graph. In this paper we study the threshold for the existence of a giant component in two expanding graph processes.

Although the analysis of our processes was motivated by attempts to model the internet, we present the results for their intrinsic interest. Indeed other time-dependent random processes will obviously provide better models of the web graph, yielding e.g. a degree sequence like that of the web graph (see Aiello Chung and Lu 2002, Barabási Albert and Jeong 1999).

## 2 The Models

We are interested in the following random process UGROW with parameter a constant $p$, with $0 < p \leq 1$, for constructing an undirected graph.

0. Initialize with the single vertex 1.

1. For $i = 2, ..., n$ add vertex $i$ and with probability $p$ add an edge between $i$ and a vertex chosen uniformly at random from $1, ..., i-1$.

We are interested in the following random process DGROW for constructing a directed graph. Again the parameters $0 \leq p_{down}, p_{extra} \leq 1$ are constants, and all choices are independent.

0. Initialize with the single vertex 1.
1. For $i = 2, ..., n$ add vertex $i$ and with probability $p_{down}$ add an arc from $i$ to a vertex chosen uniformly at random from $1, ..., i-1$.
2. For each ordered pair $(i, j)$ of vertices, add an arc from $i$ to $j$ with probability $\frac{p_{extra}}{n-1}$.

# 3 The Results

Obviously, if $p = 1$ in UGROW then the algorithm produces a spanning tree of $G$. We prove:

**Theorem 3.1** *Let $M_n$ be the maximum order (number of nodes) of a component of the $n$-node graph constructed by UGROW. Then the expected value of $M_n$ satisfies $\mathbf{E}(M_n) = \Theta(n^p)$; and for any $\epsilon > 0$ there are positive constants $c_1$ and $c_2$ such that*

$$\mathbf{P}(c_1 n^p \leq M_n \leq c_2 n^p) > 1 - \varepsilon$$

*for all $n$.*

In the directed case, we are interested in whether or not there is a 'giant' strong component, that is one with $\Omega(n)$ vertices.

**Theorem 3.2** *If $p_{down} + p_{extra} \leq 1$ then the digraph constructed by DGROW almost surely has no giant strong component.*

**Theorem 3.3** *If $p_{down} + p_{extra} > 1$ then the digraph constructed by DGROW almost surely has a giant strong component.*

Theorem 3.1 is a consequence of much finer results on the output of UGROW. We discuss these results in the next section and then turn to the directed case. We close the paper with some concluding remarks.

# 4 Analyzing UGROW

We orient each edge of the random graph we obtain to point to its endpoint of smaller index. The directed graph obtained is a random forest, consisting of a number of trees which is distributed like $1 + B(n-1, 1-p)$, where $B(n, p)$ denotes a binomial random variable with parameters $n$ and $p$. Let $N_i$ denote the order (number of vertices) of the subtree in the forest rooted at node $i$, and let $M_n = \max(N_1, \ldots, N_n)$ be the maximal order of a subtree. We will show the following.

**Lemma 4.1** *For fixed $k$,*

$$\frac{N_k}{n^p} \to \mathcal{Z}(k,p)$$

*in distribution, and $\mathcal{Z}(k,p)$ is a random variable with $\ell$-th moment*

$$\frac{\Gamma(\ell+1)\Gamma(k)}{\Gamma(k+p\ell)} .$$

Note that for $p=1$, these are the moments of the beta $(1,k-1)$ distribution when $k>1$. For $k=1$, $\mathcal{Z}(k,1)\equiv 1$.

**Lemma 4.2** *For all $\ell \geq 0$, and all $1 \leq k \leq n$,*

$$\mathbf{E}\{N_k(N_k+1)\cdots(N_k+\ell)\} \leq (\ell+1)!\left(\frac{n}{k}\right)^{p(\ell+1)} e^{p(\ell+1)/k} .$$

**Lemma 4.3** *For $t>0$,*

$$\mathbf{P}\left\{\frac{M_n}{n^p} \geq t\right\} \leq \frac{\Gamma(2+2/p)e^2\pi^2/6}{t^{2/p}} .$$

Note that Lemma 4.3 may be generalized to bounds of the form $C(a,p)/t^a$ for any $a>0$ and some constants $C(a,p)>0$. The order $n^p$ for $M_n$ is actually achieved in all cases in a probabilistic sense: for all $t>0$, we have,

$$\mathbf{P}\left\{\frac{M_n}{n^p} \leq t\right\} \leq \mathbf{P}\left\{\frac{N_1}{n^p} \leq t\right\} = \mathbf{P}\{\mathcal{Z}(k,p) \leq t\} + o(1)$$

But $\mathbf{P}\{\mathcal{Z}(k,p) \leq t\}$ tends to zero as $t \downarrow 0$:

**Lemma 4.4** *For all $p \in (0,1)$, and all $k \geq 1$, $\mathcal{Z}(k,p)$ is a continuous random variable. In particular,*

$$\lim_{t\downarrow 0} \mathbf{P}\{\mathcal{Z}(k,p) \leq t\} = 0 .$$

The forest we are studying is somewhat related to uniform random recursive trees. A uniform random recursive tree (or URRT) on $n$ nodes is a tree recursively constructed by letting the $i$-th node pick its parent uniformly and at random from among the first $i-1$ nodes. This corresponds to $p=1$ in our model. A uniform random recursive dag (or URRD) on $n$ nodes starts this process only at node $m+1$, so that the first $m$ nodes are roots. Furthermore, the $i$-th node picks $r$ nodes uniformly from among the first $i-1$ nodes to be its "parents", thus creating a directed acyclic graph. Na and Rapoport (1970), Moon (1974), Gastwirth (1977), Meir and Moon (1978), Najock and Heyde (1982), Dondajewski and Szymański (1982), Gastwirth and Bhattacharya (1984), Devroye (1987, 1988), Szymański (1987, 1990), Mahmoud (1992), Mahmoud and Smythe (1991), Pittel (1994), and Devroye and Lu (1995) have studied the URRT in some detail. A URRT of course is just a URRD with $m=1$. Dags model expression trees in which the symbols are the roots and the mathematical operators correspond to internal nodes. They also model PERT networks, and represent partial orders in general.

There is also a Pólya urn model view for our process. In Pólya urns (Pólya, 1931), one starts with a fixed finite number of urns, each having a given number of balls. An urn is picked with probability proportional to the size of the urn, and a ball is added to that urn. An urn in our setting is of course a tree in the forest. It was shown by Pólya and others (Defays, 1974, Athreya, 1969; for a survey, see Johnson and Kotz, 1977) that the proportions of the balls in the urns tends almost surely to a Dirichlet random vector. The urn occupancies are thus not concentrated in the sense that the proportion of balls in the first urn does not tend in probability to a constant. This lack of concentration is also apparent from the results below. In fact, the moment method proof of Lemma 4.4 is mimicked after the standard proof of the beta limit law for the proportion of balls in the first urn in Pólya's urn model. However, our limit law for each subtree size is not beta! In fact, the subtrees have sizes that are roughly $(n/k)^p$. Theorem 3.1 shows that the maximal tree size is $O(n^p)$ in probability.

PROOF OF LEMMA 4.1.

Consider the following process started at node $k$. Let $X_k = 1$, and for $j > k$, let $X_j$ denote the size of the subtree rooted at $k$ when $j$ nodes have been processed. When the $j$-th node is processed, note that that subtree grows by one with probability $pX_{j-1}/(j-1)$. Clearly, $X_n = N_k$. For fixed $\ell \geq 0$, it takes a moment to verify the following relationship for the $(\ell+1)$-st increasing factorial moment:

$$\begin{aligned}
&\mathbf{E}\left\{X_{j+1}(X_{j+1}+1)\cdots(X_{j+1}+\ell)\right\} \\
&\quad= \mathbf{E}\left\{X_j(X_j+1)\cdots(X_j+\ell)\right\} \\
&\qquad +(\ell+1)\mathbf{E}\left\{(X_j+1)\cdots(X_j+\ell)\times\frac{pX_j}{j}\right\} \\
&\quad= \mathbf{E}\left\{X_j(X_j+1)\cdots(X_j+\ell)\right\}\times\left(1+\frac{p(\ell+1)}{j}\right) .
\end{aligned}$$

¿From this, we have without further work,

$$\begin{aligned}
\mathbf{E}\left\{X_n(X_n+1)\cdots(X_n+\ell)\right\} &= (\ell+1)!\prod_{j=k+1}^{n}\left(1+\frac{p(\ell+1)}{j-1}\right) \\
&= (\ell+1)!\frac{\Gamma(n+p(\ell+1))\Gamma(k)}{\Gamma(k+p(\ell+1))\Gamma(n)} .
\end{aligned}$$

For fixed $k$ and $l$, we note that the right-hand side is asymptotic to

$$n^{p(\ell+1)}\frac{\Gamma(\ell+2)\Gamma(k)}{\Gamma(k+p(\ell+1))} .$$

Thus,

$$\lim_{n\to\infty}\frac{\mathbf{E}\left\{X_n(X_n+1)\cdots(X_n+\ell)\right\}}{n^{p(\ell+1)}} = \frac{\Gamma(\ell+2)\Gamma(k)}{\Gamma(k+p(\ell+1))} .$$

The limit of $\mathbf{E}\left\{(X_n/n^p)^{(\ell+1)}\right\}$ is identical. Carleman's condition applied to the limiting moments shows that these are the moments of a distribution that is uniquely determined by its moments. We call the limiting distribution $\mathcal{Z}(k,p)$. This proves Lemma 4.1.

Proof of lemma 4.2.

From the proof of Lemma 4.1, we recall

$$\begin{aligned}
&\mathbf{E}\{X_n(X_n+1)\cdots(X_n+\ell)\} \\
&\quad = (\ell+1)!\prod_{j=k+1}^{n}\left(1+\frac{p(\ell+1)}{j-1}\right) \\
&\quad \le (\ell+1)!\exp\left(\sum_{j=k}^{n-1}\frac{p(\ell+1)}{j}\right) \\
&\quad \le (\ell+1)!\exp\left(p(\ell+1)(\log(n/k)+1/k)\right) \\
&\quad \le (\ell+1)!\left(\frac{n}{k}\right)^{p(\ell+1)} e^{p(\ell+1)/k} .
\end{aligned}$$

Proof of lemma 4.3.

For $t<1$, there is nothing to prove, so assume $t \ge 1$. Let $X_n = N_k$. By Markov's inequality,

$$\begin{aligned}
\mathbf{P}\{X_n \ge tn^p\} &\le \mathbf{P}\left\{X_n(X_n+1)\cdots(X_n+\ell) \ge t^{\ell+1}n^{p(\ell+1)}\right\} \\
&\le \frac{\mathbf{E}\{X_n(X_n+1)\cdots(X_n+\ell)\}}{t^{\ell+1}n^{p(\ell+1)}} \\
&\le (\ell+1)!\left(\frac{e^{1/k}}{t^{1/p}k}\right)^{p(\ell+1)}
\end{aligned}$$

uniformly over all $n$.

In particular, if we set $\ell = \lceil 2/p \rceil - 1$, then $p(\ell+1) \ge 2$. Thus,

$$\mathbf{P}\{N_k \ge tn^p\} \le \Gamma(2+2/p)\left(\frac{e}{t^{1/p}k}\right)^2 .$$

From this, we deduce by Boole's inequality,

$$\begin{aligned}
\mathbf{P}\{\max(N_1,\ldots,N_n) \ge tn^p\} &\le \sum_{k=1}^{n}\Gamma(2+2/p)\left(\frac{e}{t^{1/p}k}\right)^2 \\
&\le \frac{\Gamma(2+2/p)e^2\pi^2/6}{t^{2/p}} .
\end{aligned}$$

Proof of lemma 4.4.

The random variable $\mathcal{Z}(k,p)$ has characteristic function given by

$$\begin{aligned}\varphi(t) &= \mathbf{E}\left\{e^{it\mathcal{Z}(k,p)}\right\} \\ &= \sum_{r=0}^{\infty} \frac{(it)^r}{r!}\mathbf{E}\left\{(\mathcal{Z}(k,p))^r\right\} \\ &= \Gamma(k)\sum_{r=0}^{\infty} \frac{(it)^r}{\Gamma(k+pr)} \\ &= \Gamma(k)\mathcal{M}_{p,k}(it)\ ,\end{aligned}$$

where $\mathcal{M}_{p,k}(z) = \sum_{r=0}^{\infty} z^r/\Gamma(k+pr)$ is the Mittag-Leffler function with parameters $p$ and $k$. This function is of semiexponential type, analytic on the positive complex halfspace (Henrici, 1986, p. 333), and thus, $\mathcal{M}_{p,k}(z) \to 0$ if $z \to \infty$ along the imaginary axis. Thus, $|\varphi(t)| \to 0$ as $|t| \to \infty$, and thus, $\mathcal{Z}(k,p)$ is a continuous random variable. As $\mathcal{Z}(k,p)$ has no atoms, it has no atom at zero, and thus, $\mathbf{P}\{\mathcal{Z}(k,p) \le t\} = o(1)$ as $t \downarrow 0$.

# 5 Analyzing DGROW

We let $D = D(n, p_{down}, p_{extra})$ be the random digraph constructed by DGROW. For each vertex $v$, we let $F(v)$ be the set of vertices which can be reached by a directed path From $v$ in $D$. We let $T(v)$ be the set of vertices from which there is a directed path To $v$ in $D$. We note that the strong component containing $v$ is exactly $T(v) \cap F(v)$. Our approach is to model the construction of $F(v)$ for each vertex using a branching process.

The expected number of arcs out of vertex $i$ in $D$ is essentially $p_{down} + p_{extra}$. (More precisely, for $i > 1$ this expected value is $p_{down} + p_{extra} - \frac{p_{down}p_{extra}}{n-1}$ whilst for $i = 1$ it is $p_{extra}$). If this value is at most 1 then it is not hard to show that almost surely the maximum size of a strong component is $o(n)$, as we now see.

**Proof of Theorem 3.2** The outdegree of a node is stochastically at most the sum of independent random variables $B(n-1, \frac{p_{extra}}{n-1})$ and $B(1, p_{down})$. Let $X_n, X_n^{(1)}, X_n^{(2)}, \ldots$ be independent random variables with this distribution. Note that $X_n$ is a sum of $n$ Poisson trials, and $\mathbf{E}(X_n) = p_{down} + p_{extra}$. We consider a Galton-Watson branching process in which the family sizes are distributed like $X_n$. Let $R = R_n$ be the random tree constructed by this process. Clearly, for any node $v$, $\mathbf{P}(|F(v)| \ge k) \le \mathbf{P}(|R| \ge k)$.

Consider first the case when $p_{down} + p_{extra} = 1 - \epsilon$ for some $\epsilon > 0$, so that the expected number of offspring in our branching process is $1 - \epsilon$. Now

$$\begin{aligned}\mathbf{P}(|R| > k) &= \mathbf{P}(\sum_{i=1}^{j}(X_n^{(i)} - 1) \ge 0 \quad \forall j = 1, \ldots, k) \\ &\le \mathbf{P}(\sum_{i=1}^{k} X_n^{(i)} \ge k).\end{aligned}$$

But $\sum_{i=1}^{k} X_n^{(i)}$ is a sum of $nk$ Poisson trials with (total) mean $\mu = (p_{down} + p_{extra})k = (1-\epsilon)k$. Hence

$$\mathbf{P}(|R| > k) \leq \mathbf{P}(\sum_{i=1}^{k} X_n^{(i)} \geq (1 + \frac{\epsilon}{1-\epsilon})\mu) \leq e^{-\frac{\epsilon^2}{2}k}$$

by standard bounds. But this last term is $o(1/n)$ for $k \geq (3/\epsilon^2)\log n$, and so in this case each component of $D$ almost surely has $O(\log n)$ nodes.

Now consider the case $p_{down} + p_{extra} = 1$, when the expected number of offspring in our process equals 1. We need to be a little more careful. Note first that, if $v$ is in a strong component of $D$ with at least $k$ nodes then $|F(v)| \geq k$. Thus,

$$\begin{aligned} & \mathbf{P}(\textit{some strong component has } \geq k \textit{ nodes}) \\ \leq\ & \mathbf{E}(\#\ \textit{of strong components with } \geq k \textit{ nodes}) \\ \leq\ & \frac{1}{k}\, \mathbf{E}(\#\ \textit{of nodes in strong components with } \geq k \textit{ nodes}) \\ \leq\ & \frac{n}{k}\, \mathbf{P}(|R| \geq k). \end{aligned}$$

We may assume that $p_{down} < 1$, since otherwise $p_{extra} = 0$ and $D$ has only trivial strong components. Note that

$$var(X_n - 1) = p_{extra}(1 - \frac{p_{extra}}{n-1}) + p_{down}(1 - p_{down}) \to 1 - p_{down}^2 > 0$$

as $n \to \infty$, and $\mathbf{E}(|X_n - 1|^3) = O(1)$. Hence by the Berry-Esseen theorem, there is a constant $c$ such that for all $n$ and $k$ we have

$$\mathbf{P}(X_n^{(1)} + \cdots + X_n^{(k)} = k-1) \leq ck^{-\frac{1}{2}}.$$

It follows (see Dwass (1969)) that, for the tree $R$ corresponding to the $X_n$ distribution, we have

$$\mathbf{P}(|R| = k) = \mathbf{P}(X_n^{(1)} + \cdots + X_n^{(k)} = k-1)/k \leq ck^{-\frac{3}{2}},$$

and so $\mathbf{P}(|R| \geq k) = O(k^{-\frac{1}{2}})$. Thus

$$\mathbf{P}(\textit{some strong component has } \geq k \textit{ nodes}) \leq \frac{n}{k}\, \mathbf{P}(|R| \geq k) = O(nk^{-\frac{3}{2}}),$$

and this last term is $o(1)$ if $k = \omega(n)n^{\frac{2}{3}}$. Thus each component of $D$ almost surely has $O(\omega(n)n^{\frac{2}{3}})$ vertices. This completes the proof of Theorem 3.2.

Conversely, if $p_{down} + p_{extra} > 1$ then we almost surely have a giant strong component, as we now show.

**Proof of Theorem 3.3**

For our branching process analysis to work, we need a final 'post-processing' stage. We will reserve a constant proportion of the extra arcs to be added in this stage. That is, for some constant $p_{final} > 0$, we add an arc from $i$ to $j$ with probability $\frac{(p_{extra} - p_{final})(1 - \frac{p_{final}}{n-1})^{-1}}{n-1}$ in step 2 (which completes the first stage), and then with probability $\frac{p_{final}}{n-1}$ in the new final stage. We use our branching process analysis to show that before this final stage we have:

**Proposition 5.1** *For some $\epsilon = \epsilon(p_{down}, p_{extra}) > 0$ there are almost surely at least $\epsilon n$ vertices in the set $A_\epsilon = \{v : |F(v)| \geq \epsilon n\}$.*

**Proposition 5.2** *For some $\delta = \delta(p_{down}, p_{extra}) > 0$ there are almost surely at least $\delta n$ vertices in the set $B_\delta = \{v : |T(v)| \geq \delta n\}$.*

It is an easy matter to show that

**Proposition 5.3** *For any $\delta, \epsilon > 0$, almost surely for every $u \in A_\epsilon$ and $v \in B_\delta$ there are at least $\frac{\delta \epsilon p_{final}^2 n}{2}$ vertices $w$ for which we add both an arc from $F(u)$ to $w$ and an arc from $w$ to $T(v)$ in the final stage.*

**Proposition 5.4** *If $|A_\epsilon||B_\delta| > n \log n$ holds for some $\delta, \epsilon > 0$, then almost surely there is an arc $xy$ with $x \in B_\delta$ and $y \in A_\epsilon$.*

Combining these last two results we see that if $|A_\epsilon||B_\delta| > n \log n$ holds for some $\delta, \epsilon > 0$, then almost surely there is an $x \in B_\delta$ such that the strong component containing $x$ has at least $\frac{\delta \epsilon p_{final}^2 n}{2}$ vertices. So to prove the theorem we need only prove Propositions 5.1 and 5.2.

Now, since the sum of the sizes of the $F(v)$ equals the sum of the sizes of the $T(v)$, if Proposition 5.1 holds for some $\epsilon > 0$ then an easy averaging argument shows that Proposition 5.2 holds for $\delta = \frac{\epsilon^2}{2}$. So, in fact we need only prove Proposition 5.1.

Before doing so, we specify our choice of $p_{final}$. We recall that in step 2, instead of adding an arc from $i$ to $j$ with probability $\frac{p_{extra}}{n-1}$, we add the arc with probability $\frac{p'_{extra}}{n-1}$ for $p'_{extra} = (p_{extra} - p_{final})(1 - \frac{p_{final}}{n-1})^{-1}$. Now, no matter how small we make $p_{final}$, Propositions 5.3 and 5.4 will still hold so by decreasing $p_{final}$ we can make $p'_{extra}$ arbitrarily close to $p_{extra}$. In particular, we want to ensure that $p_{down} + p'_{extra} > 1$. It turns out that choosing $p_{final} = \frac{p_{down} + p_{extra} - 1}{2}$ ensures this is true.

Thus, the expected number of arcs out of a vertex in step 2 exceeds 1. From now on then, we may ignore $p_{final}$ and the final stage, and just assume that $p_{down} + p_{extra} > 1$. It remains to prove (the cleaned-up version of) Proposition 5.1. However, if we try to analyze growing $F(v)$ using a simple branching process we soon run into difficulties because the 'down' arcs make it highly likely we pick vertices with low indices and so the expected outdegree of a low index vertex outside of the already picked vertices rapidly becomes less than 1.

Instead, we will think of a step in the branching process as consisting of starting with a vertex $i$ , exposing all the 'extra' arcs out of $i$ and then exposing the set of vertices reachable from these vertices by 'down' arcs. Now clearly, the expected number of vertices reachable from $j$ by down arcs is essentially $1 + p_{down} + p_{down}^2 + p_{down}^3 \cdots = \frac{1}{1-p_{down}}$ (this isn't quite true if $j$ is small e.g. if $j = 1$ this value will be 0 but if e.g. $j > \sqrt{n}$ then this value will be $\frac{1}{1-p_{down}} - o(1)$). So, the expected number of vertices reachable from $i$ in a step is essentially $\frac{p_{extra}}{1-p_{down}}$, which exceeds one.

In order to avoid the complications due to low index vertices, we actually only consider arcs which go to vertices of reasonably high index. Furthermore, we only consider arcs from $i$ added in Step 1 which go to vertices whose index is reasonably high in terms of $i$. Forthwith the details.

For a given $(p_{extra}, p_{down})$, we choose $\epsilon_1, \epsilon_2 > 0$ and $C \geq 1$ so that setting $p^* = \sum_{i=0}^{C}((1-\epsilon_2)p_{down})^i$ we have:

$$(1-\epsilon_1)p_{extra}\ p^* > 1.$$

This is possible since the inequality holds if $\epsilon_1 = \epsilon_2 = 0$ and $C = \infty$ (in which case we have $p^* = (1-p_{down})^{-1}$), and we are free to choose the $\epsilon$'s as small as we like and $C$ as large as we like.

We set $\epsilon_3 = \epsilon_1 - \frac{\epsilon_2}{2}$ and $\epsilon_4 = \epsilon_3(\frac{\epsilon_2}{2})^{C+2}$. We will restrict our attention to the subgraph $D'$ of $D$ consisting of those arcs $(i,j)$ with $j > \epsilon_4 n$.

To begin, we obtain for each node $v$, a lower bound on the size of the random set $F'(v)$, consisting of those vertices which can be reached from $v$ along a path $P$ which satisfies:

(a) for any 'extra' arc (i,j) of $P$ added in Step 2 we have $j > \frac{\epsilon_3 n}{2}$,

(b) for any 'down' arc (i,j) of $P$ added in Step 1 we have $j > \frac{\epsilon_2 i}{2}$, and

(c) any set of $C+1$ consecutive arcs of $P$ contains at least one which was added in Step 2.

We will grow $F'(v)$ iteratively. In each iteration we will explore from some vertex $i$ in $F'(v)$ by exposing all the extra arcs out of $i$ which satisfy (a) and go to new vertices, and then exposing the set of new vertices reachable from these vertices by paths of up to $C$ down arcs which satisfy (b). We begin with $F'(v) = \{v\}$, and continue until either there are no unexplored vertices of $F'(v)$ or $|F'(v)| \geq \epsilon_4 n$. Thus throughout the process, there are at most $(\epsilon_3 + \epsilon_4)n \leq \epsilon_1 n$ vertices which are either already known to be in $F'(v)$ or which have indices less than $\epsilon_3$n. In the same vein, from any vertex $i$, there are at most $\frac{\epsilon_2 i}{2} + \epsilon_4 n$ vertices which have indices less than $\frac{\epsilon_2 i}{2}$ or are already in $F'(v)$. If $i \geq (\frac{\epsilon_2}{2})^C \epsilon_3 n$ then this is less than $\epsilon_2 i$.

Consider the corresponding search tree, while it contains less than $\epsilon_4 n$ nodes. The distribution of the number of (new) children of a node $v$ is stochastically at least the distribution $D_n$ defined as follows. Take a sum of $B((1-\epsilon_1)n, \frac{p_{extra}}{n-1})$ independent random variables $Y$, where each of these random variables $Y$ takes values in $\{0, 1, \ldots, C\}$ and satisfies:

$$\text{for } \ 0 \leq i \leq C, \ \ \mathbf{P}(Y \geq i) \geq ((1-\epsilon_2)p_{down})^i.$$

Further, our choices of $\epsilon_1, \epsilon_2$ and $C$ ensure that (for $n$ sufficiently large) this distribution is stochastically at least a fixed distribution $D^*$ taking a bounded set of values $\{0, 1, \ldots, b\}$ and having mean $> 1$, where we take a sum of a truncated Poisson number of independent random variables like $Y$ above.

Thus the probability that $|F'(v)| \geq \epsilon_4 n$ is at least the probability that the Galton Watson branching process with family size distribution $D^*$ constructs a tree with at least $\epsilon_4 n$ nodes.

Consider such a Galton Watson branching process. Let its generation sizes be $Z_0 = 1, Z_2, \ldots$ and let $|R|$ be the total number of descendants. We need two facts:

$$\mathbf{P}(|R| = \infty) = \epsilon_5 > 0,$$

and

$$\mathbf{P}(|R| = \infty \mid |R| \geq \omega \log n) = 1 - o(\frac{1}{n}).$$

Let $S$ be the set of nodes $v$ such that $|F'(v)| \geq \epsilon_4 n$, and let $S'$ be the set of nodes $v$ such that $|F'(v)| \geq \omega(n) \log n$. Then $S \subseteq A_{\epsilon_4}$, and from the above we have $S' \subseteq S$ a.s., and $\mathbf{E}(|S'|) \geq \epsilon_5 n$.

To complete the proof we show that $|S'|$ is concentrated around its expected value, using the second moment method. Having exposed the set $W$ of the first up to $\omega(n) \log n$ vertices of $F'(u)$ we explore, it is quite likely that for some other vertex $v$, when we expose the first up to $\omega(n) \log n$ vertices of $F'(v)$ we will not touch $W$. Thus, $\mathbf{P}(v \in S' | u \in S') = \mathbf{P}(v \in S')(1 + o(1))$, which is enough to apply the second moment method. Thus $|S'| \geq \frac{1}{2}\epsilon_5 n$ a.s., and hence $|A_\epsilon| \geq \epsilon n$ a.s., where $\epsilon = \min(\epsilon_4, \frac{1}{2}\epsilon_5)$. Thus the proof is complete.

# 6 Concluding Remarks

We could attempt to compute the probability that $D$ has no giant component when $p_{down} + p_{extra} > 1$: we believe it is exponentially small in $n$ and our technique may perhaps be pushed to yield this.

We could also imagine a random process where each vertex throws up to $k$ edges back according to probabilities $p_1, ..., p_k$.

Finally, in Kim et al. 2002, results are given for the undirected model in which at iteration $i$, we add one of the $\binom{i}{2}$ possible edges with endpoints in $\{1, .., i\}$ with probability $p$. The authors show that the threshold for having a giant component is $p = \frac{1}{8}$ and determine bounds on the size of the largest component in the subcritical case.

# References


Adamic L. and Huberman B., "Growth Dynamics of the World Wide Web," *Nature*, vol. 401, p. 131, 1999.

Aiello W. Chung F. and Lu L., "Random Evolution in Massive Graphs," *to appear in the Handbook on Massive Data Sets*, 2002.

Aiello W. Chung F. and Lu L., "A random graph model for massive graphs," *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, p. 171-180, 2000.

K. B. Athreya, "On a characteristic property of Pólya's urn," *Stud. Sci. Math. Hung.*, vol. 4, pp. 31–35, 1969.

Barabási A. Albert R. and Jeong H., "Scale-free characteristics of random networks: the topology of the world wide web," *Physics A.*, vol. 272, pp. 173-187, 1999.

Broder A. Kumar R. Maghoul F. Raghavan P. Rajagopalan S. Stata R. Tompkins A. and Wiener J. “Graph Structure in the Web,” *Computer Networks*, vol. 33, pp. 309-321, 2000.

Cooper C. and Frieze A., “A General Model for Web Graphs,” *Proceedings of ESA 2001*, 2001.

D. Defays, “Etude du comportement asymptotique de schémas d’urnes,” *Bull. Soc. Roy. Sci. Liège*, vol. 43, pp. 26–34, 1974.

L. Devroye, “Branching processes in the analysis of the heights of trees,” *Acta Informatica*, vol. 24, pp. 277–298, 1987.

L. Devroye, “Applications of the theory of records in the study of random trees,” *Acta Informatica*, vol. 26, pp. 123–130, 1988.

L. Devroye and J. Lu, “The strong convergence of maximal degrees in uniform random recursive trees and dags,” *Random Structures and Algorithms*, vol. 6, pp. 1–14, 1995.

M. Dondajewski and J. Szymański, “On the distribution of vertex-degrees in a strata of a random recursive tree,” *Bulletin de l’Académie Polonaise des Sciences, Série des Sciences Mathématiques*, vol. 30, pp. 205–209, 1982.

M. Dwass, “The total progeny in a branching process,” *Journal of Applied Probability*, vol. 6, pp. 682–686, 1969.

J. L. Gastwirth, “A probability model of a pyramid scheme,” *The American Statistician*, vol. 31, pp. 79–82, 1977.

J. L. Gastwirth and P. K. Bhattacharya, “Two probability models of pyramid or chain letter schemes demonstrating that their promotional claims are unreliable,” *Operations Research*, vol. 32, pp. 527–536, 1984.

P. Henrici, *Applied and Computational Complex Analysis Volume 2*, vol. 1, John Wiley, New York, 1986.

N. L. Johnson and S. Kotz, *Urn Models and Their Application*, John Wiley, New York, N.Y., 1977.

Kim J. Kohayakawa Y. McDiarmid C. Reed B. Spencer J. Vu V., “An expanding graph process,” *in preparation* , 2002.

Kleinberg J. Kumar S. Raghavan P. Rajagopalan S. and Tomkins A., “The web as a graph: Measurements, models and methods,” *Proceedings of the International Conference on Combinatorics and Computing*, 1999.

Kumar S. Raghavan P. Rajagopalan S. and Tomkins A., “Extracting Large-scale Knowledge Bases from the Web,” *Proceedings of the 25th VLDB conference*, 1999.

H. M. Mahmoud and R. T. Smythe, "On the distribution of leaves in rooted subtrees of recursive trees," *Annals of Applied Probability*, vol. 1, pp. 406–418, 1991.

H. M. Mahmoud, *Evolution of Random Search Trees*, John Wiley, New York, 1992.

A. Meir and J. W. Moon, "Path edge-covering constants for certain families of trees," *Utilitas Mathematica*, vol. 14, pp. 313–333, 1978.

Molloy M. and Reed B., "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms* vol. 6, pp. 161-179, 1995.

J. W. Moon, "On the maximum degree in a random tree," *Michigan Mathematical Journal*, vol. 15, pp. 429–432, 1968.

J. W. Moon, "The distance between nodes in recursive trees," *London Mathematical Society Lecture Notes*, vol. 13, pp. 125–132, Cambridge University Press, London, 1974.

H. S. Na and A. Rapoport, "Distribution of nodes of a tree by degree," *Mathematical Biosciences*, vol. 6, pp. 313–329, 1970.

D. Najock and C. C. Heyde, "On the number of terminal vertices in certain random trees with an application to stemma construction in philology," *Journal of Applied Probability*, vol. 19, pp. 675–680, 1982.

B. Pittel, "Note on the heights of random recursive trees and random $m$-ary search tres," *Random Structures and Algorithms*, vol. 5, pp. 337–347, 1994.

G. Pólya, "Sur quelques points de la théorie de probabilité," *Ann. Inst. Henri Poincaré*, vol. 1, pp. 117–161, 1931.

Strogatz S. and Watts D., "Collective Behaviour of Small World Networks," *Nature*, vol. 393, 1998.

J. Szymański, "On a nonuniform random recursive tree," *Annals of Discrete Mathematics*, vol. 33, pp. 297–306, 1987.

J. Szymański, "On the maximum degree and the height of a random recursive tree," in: *Random Graphs 87*, (edited by M. Karoński, J. Jaworski and A. Ruciński), pp. 313–324, John Wiley, Chichester, 1990.

**Luc Devroye**
McGill University
Montreal, Canada
luc@cs.mcgill.ca

**Colin McDiarmid**
Oxford University
Oxford, UK
cmcd@stats.ox.ac.uk

**Bruce Reed**
McGill University
Montreal, Canada
breed@cs.mcgill.ca