Depth-First Search performance in a random digraph with geometric degree distribution

Svante Janson and Philippe Jacquet

Aofa, Philadelphia (virtually, alas)
22 June, 2022

Background and abstract

Donald Knuth asked about properties of Depth-First Search (DFS) in a random digraph.

Background and abstract

Donald Knuth asked about properties of Depth-First Search (DFS) in a random digraph.

We give some answers.

Full paper in preparation.

Full paper in preparation.

Some results first proved using generating functions.

The model

DFS in a digraph:

- Start with an arbitrary vertex.
- Explore the arcs from that vertex one by one
- When an arc is found that leads to a vertex that has not been seen before, explore all arcs from this vertex in the same way, recursively, before proceeding
- When there are no more arcs found, we have created a tree containing all descendants of the first vertex. If there is any vertex left, start again with a new vertex, and repeat until all vertices are explored.

This generates a spanning forest (the *depth-first forest*) in the digraph.

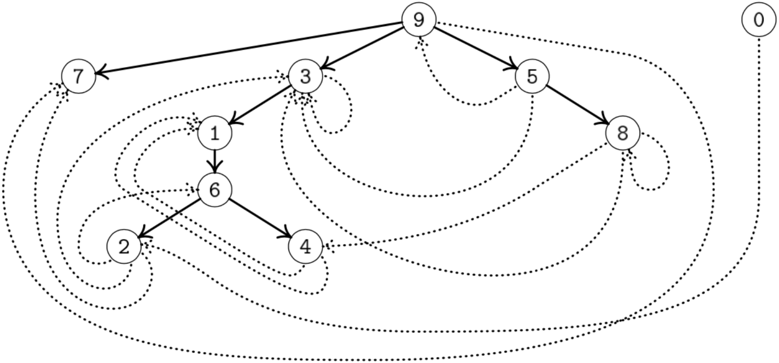

Figure: Example of a depth-first forest (solid). (Courtesy of Donald Knuth.)

The random digraph:

- The random digraph has n vertices. Each vertex v has a random outdegree η_v; these are i.i.d. (independent and identically distributed). We denote their distribution by $\mathbf{P}$.
- The endpoint of the η_v arcs from v are chosen randomly, uniformly among all vertices (including v) and independently.
- We consider asymptotics as $n \to \infty$ and the degree distribution $\mathbf{P}$ is fixed. (Extensions to $\mathbf{P} = \mathbf{P}_n$?)

Remark. The digraph is really a multi-digraph, since loops and parallel arc may occur. (But they are few, and unimportant.)

Note that we can generate the random digraph while we do the DFS: each time we visit a new vertex v, we sample its outdegree η_v.

Note that we can generate the random digraph while we do the DFS: each time we visit a new vertex v, we sample its outdegree η_v.

Knuth asked in particular about the case when **P** is a geometric distribution $\mathrm{Ge}(1-p)$ $(0 < p < 1)$. In this case, the process is almost memory-free; each time we arrive or return to a vertex, we toss a coin and:

- with probability p construct and follow a new arc to a random vertex;
- with probability $1-p$ leave the vertex and return to its parent.

Note that we can generate the random digraph while we do the DFS: each time we visit a new vertex v, we sample its outdegree η_v.

Knuth asked in particular about the case when **P** is a geometric distribution $\mathrm{Ge}(1-p)$ $(0 < p < 1)$. In this case, the process is almost memory-free; each time we arrive or return to a vertex, we toss a coin and:

- with probability p construct and follow a new arc to a random vertex;
- with probability $1-p$ leave the vertex and return to its parent.

Today mainly the geometric case. The general case is treated in detail, by a variation of the method, in our full paper (in preparation).

Geometric outdegrees

Assume that the outdegree distribution is geometric $\mathrm{Ge}(1-p)$ for some fixed $0 < p < 1$, and thus has mean

$$\lambda := \mathbb{E}\,\eta = \frac{p}{1-p}.$$

Let v_t be the t-th vertex discovered by the DFS ($t = 1, \ldots, n$), and let $d(t)$ be the depth of v_t in the resulting depth-first forest, i.e., the number of tree edges that connect the root of the current tree to v_t. The first found vertex v_1 is a root, and thus $d(1) = 0$.

Geometric outdegrees

Assume that the outdegree distribution is geometric $\mathrm{Ge}(1-p)$ for some fixed $0 < p < 1$, and thus has mean

$$\lambda := \mathbb{E}\,\eta = \frac{p}{1-p}.$$

Let v_t be the t-th vertex discovered by the DFS ($t = 1, \dots, n$), and let $d(t)$ be the depth of v_t in the resulting depth-first forest, i.e., the number of tree edges that connect the root of the current tree to v_t. The first found vertex v_1 is a root, and thus $d(1) = 0$.

The quantity $d(t)$ follows a Markov chain with transitions ($1 \leq t < n$):

- $d(t+1) = d(t) + 1$. For some $k \geq 1$, v_t has at least k outgoing arcs, the first $k-1$ arcs lead to vertices already visited, and the kth arc leads to a new vertex. Probability:

$$\sum_{k=1}^{\infty} p^k \left(\frac{t}{n}\right)^{k-1} \left(1 - \frac{t}{n}\right) = \frac{(1-t/n)p}{1-pt/n}. \tag{1}$$

- $d(t+1) = d(t)$, assuming $d(t) > 0$. All arcs from v_t lead to already visited vertices, i.e., the first case does not happen; furthermore, the parent of v_t has at least one future (not yet seen) arc leading to an unvisited vertex. These two events are independent. Moreover, by the lack-of-memory property, the second event has the probability (1). Probability:

$$\left(1 - \frac{(1-t/n)p}{1-pt/n}\right) \frac{(1-t/n)p}{1-pt/n}. \tag{2}$$

- $d(t+1) = d(t) - \ell$, assuming $d(t) > \ell \geq 1$.
  Similar. All arcs from v_t lead to already visited vertices, and so do all future arcs from the ℓ nearest ancestors of v_t, but not from the $(\ell+1)$th. Probability:

$$\left(1 - \frac{(1-t/n)p}{1-pt/n}\right)^{\ell+1} \frac{(1-t/n)p}{1-pt/n}. \tag{3}$$

- $d(t+1) = d(t) - \ell$, assuming $d(t) = \ell \geq 0$.
  Similar, except that the $(\ell+1)$th ancestor does not exist and we ignore it. Probability:

$$\left(1 - \frac{(1-t/n)p}{1-pt/n}\right)^{\ell+1}. \tag{4}$$

We can summarize this in the formula

$$d(t+1) = \max\bigl(d(t) + 1 - \xi_t, 0\bigr)$$

where ξ_t is a random variable, independent of the history, with the geometric distribution $\mathrm{Ge}(\pi_t)$, where

$$\pi_t := \frac{(1 - t/n)p}{1 - pt/n} = 1 - \frac{1-p}{1-pt/n}.$$

Define

$$\widetilde{d}(t) := \sum_{i=1}^{t-1} (1 - \xi_i),$$

this is a sum of independent random variables. Induction yields

$$d(t) = \widetilde{d}(t) - \min_{1 \le j \le t} \widetilde{d}(j), \qquad 1 \le t \le n.$$

Remark. Similar formulas have been used for other, related, problems with random graphs and trees, where trees have been coded as walks, see for example Aldous (1997). Note that in our case, $\widetilde{d}(t)$ may have negative jumps of arbitrary size.

Let $\theta := t/n$. Then, uniformly for $0 \le \theta \le \theta^*$ for any $\theta^* < 1$,

$$\begin{aligned}\mathbb{E}\left[\widetilde{d}(t)\right] &= \sum_{i=1}^{t-1}(1 - \mathbb{E}\,\xi_i) = \sum_{i=1}^{t-1}\left(1 - \frac{1-p}{p(1-i/n)}\right) \\ &= n\widetilde{\ell}(\theta) + O(1),\end{aligned}$$

where

$$\widetilde{\ell}(\theta) := \int_0^\theta \left(1 - \frac{1}{\lambda(1-x)}\right) \mathrm{d}x = \theta + \frac{1}{\lambda}\log(1-\theta).$$

Note that the derivative $\widetilde{\ell}'(\theta) = 1 - \lambda^{-1}/(1-\theta)$ is (strictly) decreasing on $(0,1)$, i.e., $\widetilde{\ell}$ is concave, and $\widetilde{\ell}(\theta) \to -\infty$ as $\theta \to 1$.

Two (three) cases

- If $\lambda > 1$ (*the supercritical case*), then $\widetilde{\ell}'(0) > 0$. There exists $\theta_1 \in (0,1)$ with $\widetilde{\ell}(\theta_1) = 0$.
  $\widetilde{\ell}(\theta) > 0$ for $\theta \in (0, \theta_1)$, and $\widetilde{\ell}(\theta) < 0$ for $\theta \in (\theta_1, 1)$.
- If $\lambda = 1$ (*critical*) or $\lambda < 1$ (*subcritical*), then $\widetilde{\ell}'(0) \leq 0$.
  $\widetilde{\ell}(\theta) < 0$ for $\theta \in (0,1)$. Let $\theta_1 := 0$.

In all cases, θ_1 is the largest solution in $[0,1]$ to

$$\log(1 - \theta_1) = -\lambda\theta_1.$$

or

$$1 - \theta_1 = \exp(-\lambda\theta_1)$$

which shows that θ_1 equals the survival probability of a Galton–Watson process with $\mathrm{Po}(\lambda)$ offspring distribution.

Let $\widetilde{\ell}^{+}(\theta) := max\big(\widetilde{\ell}(\theta), 0\big)$.

Theorem

$$\max_{1\leq t\leq n}\big|d(t) - n\widetilde{\ell}^{+}(t/n)\big| = O_{L^2}(n^{1/2}).$$

Let $\widetilde{\ell}^{+}(\theta) := max\big(\widetilde{\ell}(\theta), 0\big)$.

Theorem

$$\max_{1\le t\le n}\big|d(t) - n\widetilde{\ell}^{+}(t/n)\big| = O_{L^2}(n^{1/2}).$$

In the supercritical case, it follows that the depth-first forest whp consist of:

- Possibly one or a few small trees for small t
- One giant tree of size $\approx \theta_1 n$
- linearly many small trees for $t > n\theta_1$

Moreover, Gaussian fluctuations of $\widetilde{d}(t)$ and $d(t)$:

Theorem

In the supercritical case, in the space $D[0, \theta_1)$,

$$n^{-1/2}\big(d(\lfloor n\theta \rfloor) - n\widetilde{\ell}^{+}(\theta)\big) \xrightarrow{\mathrm{d}} Z(\theta)$$

where $Z(\theta)$ is the continuous Gaussian process

$$Z(\theta) = B\left(\lambda^{-2}\frac{\theta}{1-\theta} - \lambda^{-1}\log(1-\theta)\right)$$

for a Brownian motion $B(x)$.

Let θ_0 be the maximum point of $\widetilde{\ell}^{+}(\theta)$.
If $\lambda > 1$ the $\theta_0 = 1 - \lambda^{-1}$, otherwise $\theta_0 = 0$.

Corollary

The height Υ of the depth-first forest is

$$\Upsilon := \max_{1 \le t \le n} d(t) = \upsilon n + O_{L^2}(n^{1/2}),$$

where

$$\upsilon = \upsilon(p) := \widetilde{\ell}^{+}(\theta_0) = \begin{cases} 0, & 0 < \lambda \le 1, \\ 1 - \lambda^{-1} - \lambda^{-1} \log \lambda, & \lambda > 1. \end{cases}$$

Let θ_0 be the maximum point of $\widetilde{\ell}^+(\theta)$.
If $\lambda > 1$ the $\theta_0 = 1 - \lambda^{-1}$, otherwise $\theta_0 = 0$.

Corollary

The height Υ of the depth-first forest is

$$\Upsilon := \max_{1 \le t \le n} d(t) = \upsilon n + O_{L^2}(n^{1/2}),$$

where

$$\upsilon = \upsilon(p) := \widetilde{\ell}^+(\theta_0) = \begin{cases} 0, & 0 < \lambda \le 1, \\ 1 - \lambda^{-1} - \lambda^{-1} \log \lambda, & \lambda > 1. \end{cases}$$

Moreover, the height Υ is asymptotically normally distributed.

Corollary

The average depth $\overline{d}$ in the depth-first forest is

$$\overline{d} := \frac{1}{n}\sum_{t=1}^{n} d(t) = \alpha n + O_{L^2}(n^{1/2}),$$

where

$$\alpha = \alpha(p) := \frac{1}{2}\theta_1^2 - \frac{1}{\lambda}\Big((1-\theta_1)\log(1-\theta_1)+\theta_1\Big) = \frac{\lambda-1}{\lambda}\theta_1 - \frac{1}{2}\theta_1^2.$$

We have $\alpha = 0$ if and only if $\lambda \le 1$, i.e., $p \le 1/2$.

Corollary

The average depth $\overline{d}$ in the depth-first forest is

$$\overline{d} := \frac{1}{n}\sum_{t=1}^{n} d(t) = \alpha n + O_{L^2}(n^{1/2}),$$

where

$$\alpha = \alpha(p) := \frac{1}{2}\theta_1^2 - \frac{1}{\lambda}\Big((1-\theta_1)\log(1-\theta_1)+\theta_1\Big) = \frac{\lambda-1}{\lambda}\theta_1 - \frac{1}{2}\theta_1^2.$$

We have $\alpha = 0$ if and only if $\lambda \leq 1$, i.e., $p \leq 1/2$.

Remark. When $p > \frac{1}{2}$, the height is thus linear in n, unlike many other types of random trees.

Types of arcs

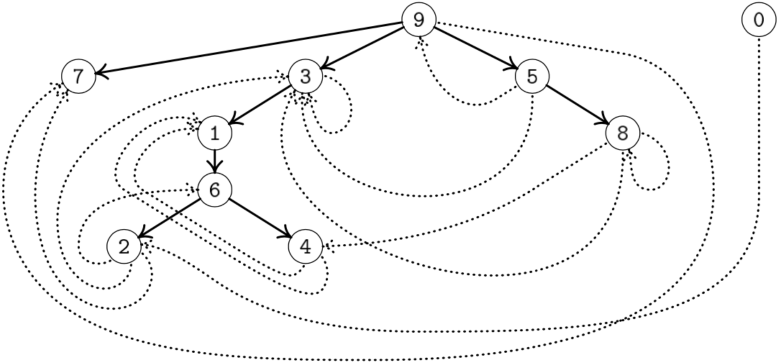

Figure: Example of a depth-first forest (jungle). (Courtesy of Donald Knuth.) Tree arcs are solid (*e.g.* ⑨→ ③). For example, ③--→ ③ is a loop, ②--→ ③ is a back arc, ⑨--→ ⑦ is a forward arc, ⑧--→ ④ and ⓪--→ ② are cross arcs.

Theorem

Let L, T, B, F and C be the numbers of loops, tree arcs, back arcs, forward arcs, and cross arcs in the random digraph. Then

$$\begin{aligned} L &= O_{L^2}(1), \\ T &= \tau n + O_{L^2}(n^{1/2}), \\ B &= \beta n + O_{L^2}(n^{1/2}), \\ F &= \varphi n + O_{L^2}(n^{1/2}), \\ C &= \chi n + O_{L^2}(n^{1/2}), \end{aligned}$$

where

$$\begin{aligned} \tau := \chi &:= \theta_1 + \frac{\lambda}{2}(1-\theta_1)^2, \\ \beta := \varphi &:= \lambda\alpha = (\lambda - 1)\theta_1 - \frac{\lambda}{2}\theta_1^2. \end{aligned}$$

The equalitites $\tau = \chi$ and $\beta = \varphi$ mean asymptotic equality of the corresponding expectations of numbers of arcs. In fact, there are exact equalities.

Theorem

For any n, $\mathbb{E}\, T = \mathbb{E}\, C$ and

$$\mathbb{E}\, B = \mathbb{E}\, F = \lambda\, \mathbb{E}\, \overline{d} = \beta n + O(n^{1/2}).$$

Remark. Knuth conjectures, based on exact formulas for small n, that, much more strongly, B and F have the same distribution for every n. (Note that T and C do not have the same distribution; we have $T \leq n - 1$, while C may take arbitrarily large values.)

General outdegree distribution **P**

For a general outdegree distribution, the depth is no longer a Markov chain.

Substitute: The DFS uses a stack of unexplored arcs, for which we have seen the start vertex but not the end. The stack evolves as follows:

- S1 If the stack is empty, pick a new vertex v that has not been seen before (if there is no such vertex, we have finished). Otherwise, pop the last arc from the stack, and reveal its endpoint v (which is uniformly random over all vertices). If v already is seen, repeat.
- S2 (v is now a new vertex) Reveal the outdegree η of v and add to the stack η new arcs from v, with unspecified endpoints. GOTO S1

The size $I(t)$ of the stack is a Markov chain.

Let

$$\widetilde{\iota}^{+}(\theta) := \begin{cases} \lambda\theta + \log(1-\theta), & 0 \le \theta \le \theta_1, \\ 0, & \theta_1 \le \theta \le 1. \end{cases}$$

Theorem

Suppose that the outdegree distribution has finite variance. Then

$$\max_{1 \le t \le n} \left| I(t) - n\widetilde{\iota}^{+}(t/n) \right| = O_{L^2}(n^{1/2}).$$

The depth $d(t)$ and other properties can be recovered from $I(t)$.

Many (but not all) results extend to general $\mathbf{P}$.