The number of descendants in a random directed acyclic graph

Svante Janson

29th Nordic Conference in Mathematical Statistics,
NORDSTAT 2023, Gothenburg
21 June, 2023

Directed acyclic graph (dag)

A *dag* is a directed acyclic (multi)graph.

A *d-dag* is a dag where all vertices have outdegrees d, except one or several *roots* with outdegree 0.

d is a positive integer; we will take $d = 2$.

Directed acyclic graph (dag)

A *dag* is a directed acyclic (multi)graph.

A *d-dag* is a dag where all vertices have outdegrees d, except one or several *roots* with outdegree 0.

d is a positive integer; we will take $d = 2$.

It is sometimes natural to direct all edges in the opposite direction.

Our model: random d-dag

The random d-dag D_n on n vertices is constructed recursively:

1. Start with a single root 1,
2. Add vertices $2, 3, \ldots, n$ one by one. Each new vertex k is given d outgoing edges with endpoints uniformly and independently chosen at random among the already existing vertices $\{1, \ldots, k-1\}$.

(We thus allow multiple edges, so D_n is a directed multigraph.)

Remark. For $d = 1$, the model becomes the well known *random recursive tree*; this case is quite different from $d > 1$ and is excluded below.

Our model: random d-dag

The random d-dag D_n on n vertices is constructed recursively:

1. Start with a single root 1,
2. Add vertices $2, 3, \ldots, n$ one by one. Each new vertex k is given d outgoing edges with endpoints uniformly and independently chosen at random among the already existing vertices $\{1, \ldots, k-1\}$.

(We thus allow multiple edges, so D_n is a directed multigraph.)

Remark. For $d = 1$, the model becomes the well known *random recursive tree*; this case is quite different from $d > 1$ and is excluded below.

Two possible minor variations (asymptotically the same for us):

1. Start with any number $m \geq 1$ of roots.
2. Select the d parents of a new node without replacement, thus not allowing multiple edges. (Start with $\geq d$ roots.)

Earlier results

This model has been studied by several authors, mainly in computer science, for example as a model for a random circuit where each gate has d inputs chosen at random (Díaz, Serna, Spirakis & Torán 1994, and others).

Earlier results include results on vertex degrees and leaves, and on lengths of paths and depth.

Tiffany Lo, a postdoc in Uppsala working with me and Cecilia Holmgren, has just shown results on subgraphs (unpublished).

Problem today

Problem

How many descendants does vertex n have?

In other words, how many vertices can be reached by a directed path from vertex n? In the random circuit interpretation, this is the number of gates (and inputs) that are used in the calculation of a given output.

As far as we know, this problem was first considered by Knuth (2023).

Problem today

Problem

How many descendants does vertex n have?

In other words, how many vertices can be reached by a directed path from vertex n? In the random circuit interpretation, this is the number of gates (and inputs) that are used in the calculation of a given output.

As far as we know, this problem was first considered by Knuth (2023).

Notation:
D_n is the random d-dag defined above.
$\widehat{D}_n$ is the subdigraph of D_n consisting of n and all vertices and edges that can be reached by a directed path from vertex n.
$X_n := |\widehat{D}_n|$, the number of descendants of n.

We colour $\widehat{D}_n$ red.

Main result

Let χ_4 denote a random variable with the $\chi(4)$ distribution.

Theorem

Let $d = 2$. Then, as $n \to \infty$,

$$X_n/\sqrt{n} \xrightarrow{\mathrm{d}} \frac{\pi}{2\sqrt{2}}\chi_4$$

with convergence of all moments. Hence, for every fixed $r > 0$,

$$\mathbb{E}\, X_n^r \sim \left(\frac{\pi}{2}\right)^r \Gamma\left(\frac{r}{2}+2\right) n^{r/2}$$

and, in particular,

$$\mathbb{E}\, X_n \sim \frac{3\pi^{3/2}}{8}\sqrt{n}.$$

Analysis

We construct the red subgraph $\widehat{D}_n$ backwards, going from vertex n backwards to 1:

1. Start by declaring vertex n to be *red*, and all others *black*. Let $k := n$.
2. If vertex k is red, then create two new edges from that vertex, with endpoints that are randomly drawn from $1, \ldots, k-1$, and declare these endpoints red.
   If k is black, delete k (and do nothing else).
3. If $k = 2$ then STOP; otherwise let $k := k - 1$ and REPEAT from 2.

Definitions. For $k = n-1, \ldots, 1$:

Y_k is the number of red edges that start in $\geq k+1$ and end in $\leq k$.

Z_k is the number of these edges that end in k.

$J_k := \mathbf{1}\{Z_k \geq 1\}$, which equals the indicator that k is red (i.e. can be reached from n).

$$Y_{k-1} = Y_k - Z_k + 2J_k = Y_k - Z_k + 2 \cdot \mathbf{1}\{Z_k \geq 1\}.$$

Definitions. For $k = n-1, \dots, 1$:

Y_k is the number of red edges that start in $\geq k+1$ and end in $\leq k$.

Z_k is the number of these edges that end in k.

$J_k := \mathbf{1}\{Z_k \geq 1\}$, which equals the indicator that k is red (i.e. can be reached from n).

$$Y_{k-1} = Y_k - Z_k + 2J_k = Y_k - Z_k + 2 \cdot \mathbf{1}\{Z_k \geq 1\}.$$

Now, do not reveal the endpoint of the edges until needed. Then $Y_{n-1}, \dots, Y_1$ is a Markov chain. Conditioned on the history, Z_k has a binomial distribution

$$Z_k \in \mathrm{Bin}(Y_k, 1/k).$$

Thus we have a stochastic recursion of Markov type for Y_k.

Simple calculations yield ($\mathcal{F}_k$ is the σ-field generated by the history)

$$\mathbb{E}\big(Y_{k-1} \mid \mathcal{F}_k\big) = Y_k - \tfrac{1}{k} Y_k + 2\big(1 - (1 - \tfrac{1}{k})^{Y_k}\big).$$
$$\mathbb{E}\big(Y_{k-1} \mid \mathcal{F}_k\big) \le Y_k + \tfrac{1}{k} Y_k = \tfrac{k+1}{k} Y_k.$$

Define

$$W_k := (k+1) Y_k,$$

Then

$$\mathbb{E}\big(W_{k-1} \mid \mathcal{F}_k\big) = k\,\mathbb{E}\big(Y_{k-1} \mid \mathcal{F}_k\big) \le (k+1) Y_k = W_k.$$

Thus $W_0, \ldots, W_{n-1}$ is a reverse supermartingale.
(I.e. W_{-j}, $-(n-1) \le j \le 0$, is a supermartingale.)

Phase I: a Yule process

For $n > k \geq n_1 := \lfloor n/\log n \rfloor$, there are w.h.p. no collisions (two edges with the same endpoint) Thus the process is essentially a branching process, where an individual born at x lives until xU with $U \in U(0,1)$, and then splits into 2 children. (Recall that the time x goes backwards.)

Changing time to $t := \log(n/x) \in (0, \infty)$ gives a Yule process (binary splitting and Exp(1) life lengths), started with 2 individuals (edges).

Thus, for $n > k \geq n_1$, the red digraph $\widehat{D}_n$ is essentially a Yule tree.

Phase I: a Yule process

For $n > k \geq n_1 := \lfloor n / \log n \rfloor$, there are w.h.p. no collisions (two edges with the same endpoint) Thus the process is essentially a branching process, where an individual born at x lives until xU with $U \in U(0,1)$, and then splits into 2 children. (Recall that the time x goes backwards.)

Changing time to $t := \log(n/x) \in (0, \infty)$ gives a Yule process (binary splitting and Exp(1) life lengths), started with 2 individuals (edges).

Thus, for $n > k \geq n_1$, the red digraph $\widehat{D}_n$ is essentially a Yule tree. It follows that (using superscript $^{(n)}$ for clarity)

$$\Xi^{(n)} = \frac{W_{n_1}^{(n)}}{n} \xrightarrow{\mathrm{d}} \xi \in \Gamma(2).$$

Phase II: nothing happens for a long time

Fix $n_2 = n_2^{(n)}$ such that $n_1 > n_2 \gg \sqrt{n}$.

Simple estimates of drift and variance for the supermartingale W_k shows that W_k/n is essentially constant for $k \in [n_2, n_1]$. Formally

$$\max_{n_1 \geq k \geq n_2} \left| \frac{W_k}{n} - \Xi^{(n)} \right| \xrightarrow{\mathrm{p}} 0.$$

Phase III: deterministic decay from a random level

For $k \leq n_2$, the process W_k still evolves (asymptotically) in a deterministic way (this is essentially a law of large numbers), but from the random level in Phase II.

Martingale methods yield, again using simple estimates of drift and variance,

$$n^{-1} W^{(n)}_{t\sqrt{n}} \xrightarrow{\mathrm{d}} t^2 \mathcal{B}(t) \qquad \text{in } D[0,\infty)$$

where the stochastic process $\mathcal{B}(t)$ is differentiable and satisfies

$$\mathcal{B}'(t) = -\frac{2}{t}\bigl(1 - e^{-\mathcal{B}(t)}\bigr).$$

Solving this equation yields, using Phase II as an initial condition,

$$n^{-1} W^{(n)}_{t\sqrt{n}} - t^2 \log\bigl(1 + \Xi^{(n)}/t^2\bigr) \to 0 \qquad \text{in } D[0,\infty).$$

The number of red vertices

We have (recall that J_k is the indicator that k is red.)

$$\mathbb{E}(J_k \mid \mathcal{F}_k) = 1 - \left(1 - \frac{1}{k}\right)^{Y_k} = 1 - \left(1 - \frac{1}{k}\right)^{W_k/k}$$

and, again using martingale methods, it follows that

$$\frac{X^{(n)}}{\sqrt{n}} = \int_0^\infty \frac{\Xi^{(n)}}{\Xi^{(n)} + t^2}\,\mathrm{d}t + o_{\mathrm{p}}(1) = \frac{\pi}{2}\sqrt{\Xi^{(n)}} + o_{\mathrm{p}}(1)$$
$$\xrightarrow{\mathrm{d}} \frac{\pi}{2}\sqrt{\xi}, \qquad \xi \in \Gamma(2).$$

(Where $o_{\mathrm{p}}(1) \xrightarrow{\mathrm{p}} 0$.)

QED

Conclusions

Asymptotically:

1. The number of red vertices is of order $n^{1/2}$.
2. Most of these are in the range $k = O(n^{1/2})$, where the density of red vertices is positive.
3. This density is random. However, the random choices in this dense region do not matter (law of large numbers); the density (and thus the total number) is determined by the random choices for the few red vertices k of order n.

References

Svante Janson: Descendants in a random directed acyclic graph.
arXiv:2302.12467
http://www2.math.uu.se/∼svante/papers/#374

and references to other authors there