Variations on depth-first search performance in random digraphs

Philippe Jacquet, Svante Janson

▶ **To cite this version:**

Philippe Jacquet, Svante Janson. Variations on depth-first search performance in random digraphs. AofA 2022 - 33rd International Conference on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms, Jun 2022, Philadelphia, PA, United States. hal-03537124

HAL Id: hal-03537124
https://hal.science/hal-03537124

Submitted on 20 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

VARIATIONS ON DEPTH-FIRST SEARCH PERFORMANCE IN RANDOM DIGRAPHS

PHILIPPE JACQUET AND SVANTE JANSON

Abstract. We present an analysis of the depth-first search algorithm in a random digraph model with geometric outdegree distribution. This problem posed by Don Knuth in his next to appear volume of The Art of Computer Programming gives interesting insight in one of the most elegant and efficient algorithm for graph analysis due to Tarjan.

1. Introduction

The motivation of this paper is a new section in Donald Knuth's *The Art of Computer Programming* [1], which is dedicated to Depth-First Search (DFS) in a digraph. We refer to [1] for the definition of DFS as well as for historical notes. Note that the digraphs in [1] and here are multi-digraphs, where loops and multiple arcs are allowed. The DFS algorithm generates a spanning forest (the *depth-first forest*) in the digraph, with all arcs in the forest directed away from the roots. Our main purpose is to study the distribution of the depth of vertices in the depth-first forest, starting with a random digraph G.

Furthermore, the DFS algorithm in [1] classifies the arcs in the digraph into the following five types, see Figure 1 for examples:

- *loops*;
- *tree arcs*, the arcs in the resulting depth-first forest;
- *back arcs*, the arcs which point to an ancestor of the current vertex in the current tree;
- *forward arcs*, the arcs which point to an already discovered descendant of the current vertex in the current tree;
- *cross arcs*, all other arcs (these point to an already discovered vertex which is neither a descendant nor an ancestor of the current vertex, and might be in another tree).

We will discuss the numbers of arcs of different types. (See further the exercises in [1].)

The random digraph model that we consider has n vertices and a given outdegree distribution $\mathbf{P}$. The outdegrees (number of outgoing arcs) of the n vertices are independent random numbers with this distribution. The endpoint of each arc is uniformly selected at random among the n vertices, independently of all other arcs. (Therefore, an arc can loop back to the starting vertex, and multiple arcs can occur.) We consider asymptotics as $n \to \infty$ for a fixed outdegree distribution.

We will focus on the case of a geometric outdegree distribution; the lack-of-memory property in this case leads to interesting features and a simpler analysis. The paper will study the following outdegree distribution in the following order:

- a geometric distribution;
- a shifted geometric distribution (starting from integer 1 instead of zero);
- a general distribution.

Date: 11 January 2022.

Key words and phrases. Combinatorics, Depth-First Search, Random Digraphs.

Supported by the Knut and Alice Wallenberg Foundation.

We thank Donald Knuth for posing us questions and conjectures that led to the present paper.

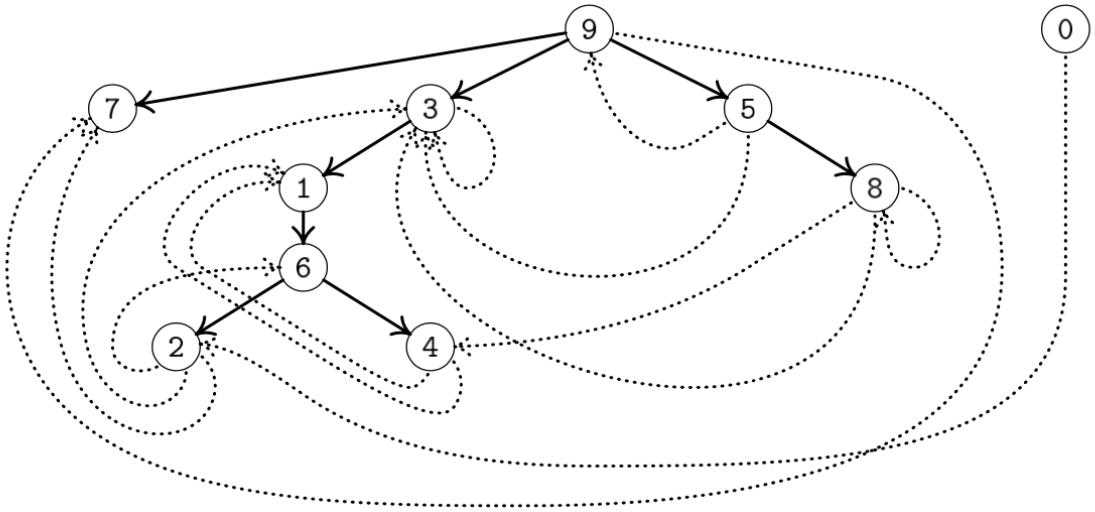

FIGURE 1. Example of a depth-first forest (jungle) from [1], by courtesy of Donald Knuth. Tree arcs are solid (*e.g.* ⑨→ ③). For example, ③⇢ ③ is a loop, ②⇢ ③ is a back arc, ⑨⇢ ⑦ is a forward arc, ⑧⇢ ④ and ⓪⇢ ② are cross arcs.

fig:forest

1.1. **Some notation.** As usual, w.h.p. means *with high probability*, *i.e.*, with probability $1-o(1)$.

We let $O_{L^2}(a_n)$, where (a_n) is a sequence of positive numbers, denote a sequence of random variables X_n such that $\mathbb{E}\bigl[|X_n/a_n|^2\bigr] = O(1)$. Note that this implies $|X_n| \le a_n\omega(n)$ w.h.p. for any sequence $\omega(n)\to\infty$.

The mean outdegree, *i.e.*, the expectation of $\mathbf{P}$, is denoted by λ.

2. Depth analysis with geometric outdegree distribution

Sgeo

In this section we assume that the outdegree distribution is geometric $\mathrm{Ge}(1-p)$ for some fixed $0<p<1$, and thus has mean $\lambda := p/(1-p)$. Note that in the DFS, when we find a new vertex v, we do not have to immediately reveal its outdegree. Instead, we only check whether there is at least one outgoing arc (probability p), and if so, we find its endpoint and explores this endpoint if it has not already been visited; eventually, we return to v, and then we check whether there is another outgoing arc (again probability p, by the lack-of-memory property of the geometric distribution), and so on.

In the following, by a *future arc* from some vertex, we mean an arc that at the current time has not yet been seen by the DFS.

2.1. **Depth Markov chain.** Our aim is to track the evolution of the search depth as a function of the number of t of discovered vertices. 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 ($1\le t<n$):

(i) $d(t+1)=d(t)+1$.
This happens if, for some $k\ge1$, 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 (which then

becomes v_{t+1}). The probability of this is

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

(ii) $d(t+1) = d(t)$, assuming $d(t) > 0$.
This holds if all arcs from v_t lead to already visited vertices, *i.e.*, (i) does not happen, and furthermore, the parent of v_t has at least one further arc leading to an unvisited vertex. These two events are independent. Moreover, by the lack-of-memory property, the number of further arcs from the parent of v_t has the same distribution $\mathrm{Ge}(1-p)$. Hence, the probability that one of these further arcs leads to an unvisited vertex equals the probability in (1). The probability of (ii) is thus

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

(iii) $d(t+1) = d(t) - \ell$, assuming $d(t) > \ell \ge 1$.
This happens if all arcs from v_t lead to already visited vertices, and so do all further arcs from the ℓ closest ancestors of v_t, while the $(\ell+1)$th ancestor has at least one further arc leading to an unvisited vertex. The argument in (ii) generalizes and shows that this has probability

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

(iv) $d(t+1) = d(t) - \ell$, assuming $d(t) = \ell \ge 0$.
By the same argument as in (ii) and (iii), except that the $(\ell+1)$th ancestor does not exist and we ignore it, we obtain the probability

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

Note that (iv) is the case when $d(t+1) = 0$ and thus v_{t+1} is the root of a new tree in the depth-first forest.

We can summarize (i)–(iv) in the formula

$$d(t+1) = \big(d(t) + 1 - \xi_t\big)^+, \tag{5}$$ dt+

where $x^+ := \max\{x, 0\}$, and ξ_t is a random variable, independent of the history, with the distribution

$$\mathbb{P}(\xi_t = k) = (1-\pi_t)^k\pi_t, \quad k \ge 0, \qquad \text{with} \qquad \pi_t := \frac{(1-t/n)p}{1-pt/n} = 1 - \frac{1-p}{1-pt/n}. \tag{6}$$ xi

In other words, ξ_t has the geometric distribution $\mathrm{Ge}(\pi_t)$. Define

$$\widetilde{d}(t) := \sum_{i=1}^{t-1}(1-\xi_i), \tag{7}$$ td

and note that (7) is a sum of independent random variables. Then (5) and induction yield

$$d(t) = \widetilde{d}(t) - \min_{1\le j\le t}\widetilde{d}(j), \qquad 1 \le t \le n. \tag{8}$$ dtd

We can also express these relations using generating functions. Let $p(t,z)$ be the probability generating function $\mathbb{E}\, z^{\xi_t}$ of ξ_t, *i.e.*,

$$p(t,z) := \frac{(1-t/n)p}{1-pt/n}\sum_{\ell\ge0}\left(1 - \frac{(1-t/n)p}{1-pt/n}\right)^{\ell} z^\ell = \frac{(1-t/n)p}{1-pt/n-(1-p)z}, \tag{9}$$ ptz

and let $f(t,z) := E[z^{d(t)}]$. We then have the identity, equivalent to (5),

$$f(t+1,z) = \mathcal{N}\big[R(t,z)f(t,z)\big] \tag{10}$$

where $R(t,z) := p(t,1/z)z$ and $\mathcal{N}$ is the operator on power series in $z^{\pm1}$:

$$\mathcal{N}g(z) = \Pi^+ g(z) + \Pi^- g(1) \tag{11}$$

where Π^+ is the operator which removes the strictly negative powers of z and Π^- is the operator which removes the non-negative powers of z. Thus we have, since $f(1,z)=1$,

$$f(t+1,z) = \mathcal{N}R(t,z)\mathcal{N}R(t-1,z)\mathcal{N}\cdots\mathcal{N}R(1,z). \tag{12}$$

2.2. **Main result for depth analysis.** Note first that (7) implies that the expectation of $\widetilde{d}(t)$ is

$$\mathbb{E}\big[\widetilde{d}(t)\big] = \sum_{i=1}^{t-1}(1-\mathbb{E}\,\xi_i) = \sum_{i=1}^{t-1}\Big(1-\frac{1-\pi_i}{\pi_i}\Big) = \sum_{i=1}^{t-1}\Big(1-\frac{1-p}{p(1-i/n)}\Big). \tag{13}$$ Etd1

Let $\theta := t/n$. We fix a $\theta^* < 1$ and obtain that, uniformly for $\theta \le \theta^*$,

$$\mathbb{E}\big[\widetilde{d}(t)\big] = \int_0^t \Big(1-\frac{1-p}{p(1-x/n)}\Big)\,\mathrm{d}x + O(1) = n\widetilde{\ell}(\theta) + O(1), \tag{14}$$ Etd

where

$$\widetilde{\ell}(\theta) := \int_0^\theta \Big(1-\frac{1-p}{p(1-\tau)}\Big)\,\mathrm{d}\tau = \theta + \frac{1-p}{p}\log(1-\theta). \tag{15}$$ tl

Note that the derivative $\widetilde{\ell}'(\theta) = 1-(1-p)/(p(1-\theta))$ is (strictly) decreasing on $(0,1)$, *i.e.*, $\widetilde{\ell}$ is concave. Moreover, if $p > \frac12$, which we call the *supercritical* case, then $\widetilde{\ell}'(0) > 0$, and (15) shows that $\widetilde{\ell}(\theta)$ is positive and increasing for $\theta < \theta_0 := (2p-1)/p$. After the maximum at θ_0, $\widetilde{\ell}(\theta)$ decreases and tends to $-\infty$ as $\theta \nearrow 1$. Hence, there exists a $0<\theta_1<1$ such that $\widetilde{\ell}(\theta_1)=0$; we then have $\widetilde{\ell}(\theta)>0$ for $0<\theta<\theta_1$ and $\widetilde{\ell}(\theta)<0$ for $\theta>\theta_1$. We will see that in this case the depth-first forest w.h.p. contains a giant tree, of order and height both linear in n, while all other trees are small.

On the other hand, if $p \le \frac12$ (the *subcritical* and *critical* cases), then $\widetilde{\ell}'(0) \le 0$ and $\widetilde{\ell}(\theta)$ is negative and decreasing for all $\theta \in (0,1)$. In this case, we define $\theta_0 := \theta_1 := 0$ and note that the properties just stated for $\widetilde{\ell}$ still hold (rather trivially). We will see that in this case w.h.p. all trees in the depth-first forest are small.

Note that in all cases, θ_1 is the largest solution in $[0,1)$ to, recalling $\lambda = p/(1-p)$,

$$\log(1-\theta_1) = -\lambda\theta_1. \tag{16}$$ gth1

Rgth1 **Remark 2.1.** The equation (16) may also be written $1-\theta_1 = \exp(-\lambda\theta_1)$, which shows that θ_1 is the survival probability of a Galton–Watson process with $\mathrm{Po}(\lambda)$ offspring distribution.

We define $\widetilde{\ell}^+(\theta) := [\widetilde{\ell}(\theta)]^+$. Thus, by (15) and the comments above,

$$\widetilde{\ell}^+(\theta) = \begin{cases} \theta + \frac{1-p}{p}\log(1-\theta), & 0 \le \theta \le \theta_1, \\ 0, & \theta_1 \le \theta \le 1. \end{cases} \tag{17}$$ tlp

We can now state one of our main results. Proofs are given in the next subsection.

T1 **Theorem 2.2.** *We have*

$$\max_{1\le t\le n}\big|d(t) - n\ell^+(t/n)\big| = O_{L^2}(n^{1/2}). \tag{18}$$ t1

CH **Corollary 2.3.** *The height Υ of the depth-first forest is*

$$\Upsilon := \max_{1\le t\le n} d(t) = \upsilon n + O_{L^2}(n^{1/2}), \tag{19}$$ gU

where

$$\upsilon = \upsilon(p) := \ell^+(\theta_0) = \begin{cases} 0, & 0 < p \le 1/2, \\ \frac{2p-1}{p} - \frac{1-p}{p}\log\frac{p}{1-p}, & 1/2 < p < 1. \end{cases} \tag{20}$$ gu

Moreover, we can show that the height Υ is asymptotically normally distributed. Details will be given in the full paper.

CA **Corollary 2.4.** *The average depth $\overline{d}$ in the depth-first forest is*

$$\overline{d} := \frac{1}{n}\sum_{t=1}^{n} d(t) = \alpha n + O_{L^2}(n^{1/2}), \tag{21}$$ ca

where

$$\alpha = \alpha(p) := \frac{1}{2}\theta_1^2 - \frac{1-p}{p}\Big((1-\theta_1)\log(1-\theta_1) + \theta_1\Big) = \frac{2p-1}{p}\theta_1 - \frac{1}{2}\theta_1^2 \tag{22}$$ ga

2.3. **Proofs.**

Proof of Theorem 2.2. Since (7) is a sum of independent random variables, $\widetilde{d}(t) - \mathbb{E}\,\widetilde{d}(t)$ $(t = 1, \dots, n)$ is a martingale, and Doob's inequality yields, for all $T \le n$,

$$\mathbb{E}\big[\max_{t\le T} |\widetilde{d}(t) - \mathbb{E}\,\widetilde{d}(t)|^2\big] \le 4\,\mathbb{E}\big[|\widetilde{d}(T) - \mathbb{E}\,\widetilde{d}(T)|^2\big] = 4\sum_{i=1}^{T-1} \mathrm{Var}(\xi_i). \tag{23}$$ emma

As above, fix $\theta^* < 1$, and assume, as we may, that $\theta^* > \theta_1$. Let $T^* := \lfloor n\theta^* \rfloor$, and consider first $t \le T^*$. For $i < T^*$, we have $\mathrm{Var}\,\xi_i = O(1)$, and thus, for $T = T^*$, the sum in (23) is $O(T^*) = O(n)$. Consequently, (23) yields

$$\max_{t\le T^*} \big|\widetilde{d}(t) - \mathbb{E}\,\widetilde{d}(t)\big| = O_{L^2}(n^{1/2}). \tag{24}$$

Hence, by (14),

$$M^* := \max_{t\le T^*} \big|\widetilde{d}(t) - n\widetilde{\ell}(t/n)\big| = O_{L^2}(n^{1/2}). \tag{25}$$ m*

For $t \le T^*$, the definition of M^* in (25) implies

$$\Big|\min_{1\le j\le t} \widetilde{d}(j) - n \min_{1\le j\le t} \widetilde{\ell}(j/n)\Big| \le M^*. \tag{26}$$ gal

Moreover, for $t/n \le \theta_1$, we have $\min_{1\le j\le t} \widetilde{\ell}(j/n) = O(1/n)$, while for $t/n \ge \theta_1$, we have $\min_{1\le j\le t} \widetilde{\ell}(j/n) = \widetilde{\ell}(t/n)$. Hence, for all $t \le T^*$,

$$\min_{1\le j\le t} \widetilde{\ell}(j/n) = \widetilde{\ell}(t/n) - \widetilde{\ell}^{+}(t/n) + O(1/n), \tag{27}$$

and thus, by (26),

$$\Big|\min_{1\le j\le t} \widetilde{d}(j) - n\widetilde{\ell}(t/n) + n\widetilde{\ell}^{+}(t/n)\Big| \le M^* + O(1/n). \tag{28}$$ ew

Finally, by (8), (25) and (28),

$$\big|d(t) - n\widetilde{\ell}^{+}(t/n)\big| \le 2M^* + O(1/n). \tag{29}$$ jb

This holds uniformly for $t \le T^*$, and thus, by (25),

$$\max_{1\le t\le T^*} \big|d(t) - n\ell^{+}(t/n)\big| = O_{L^2}(n^{1/2}). \tag{30}$$ jesp

It remains to consider $T^* < t \le n$. Then the argument above does not quite work, because $\pi_t \searrow 0$ and thus $\mathrm{Var}\,\xi_t \nearrow \infty$ as $t \nearrow n$. We therefore modify ξ_t. We define $\widehat{\pi}_t := \max\{\pi_t, \pi_{T^*}\}$; thus $\widehat{\pi}_t = \pi_t$ for $t \le T^*$ and $\widehat{\pi}_t > \pi_t$ for $t > T^*$. We may thus define independent random variables $\widehat{\xi}_t$ such that $\widehat{\xi}_t \sim \mathrm{Ge}(\widehat{\pi}_t)$ and $\widehat{\xi}_t \le \xi_t$ for all $t < n$. (Thus, $\widehat{\xi}_t = \xi_t$ for $t \le T^*$.)

In analogy with (7)–(8), we further define

$$\widehat{\widetilde{d}}(t) := \sum_{i=1}^{t-1}\bigl(1-\widehat{\xi}_i\bigr), \tag{31}$$ htd

$$\widehat{d}(t) := \widehat{\widetilde{d}}(t) - \min_{1\le j\le t}\widehat{\widetilde{d}}(j) = \max_{1\le j\le t}\sum_{i=j}^{t-1}\bigl(1-\widehat{\xi}_i\bigr). \tag{32}$$ hd

Since $\widehat{\xi}_i \le \xi_i$, (32) implies that $\widehat{d}(t) \ge d(t)$ for all t.

We have $\operatorname{Var}\bigl[\widehat{\xi}_t\bigr] = O(1)$, uniformly for all $t < n$, and thus the argument above yields

$$\max_{1\le t\le n}\bigl|\widehat{d}(t) - n[\widehat{\widetilde{\ell}}(t/n)]^+\bigr| = O_{L^2}(n^{1/2}), \tag{33}$$ kasp

where

$$\widehat{\widetilde{\ell}}(\theta) := \int_0^\theta \min\Bigl\{\Bigl(1-\frac{1-p}{p(1-\tau)}\Bigr),\Bigl(1-\frac{1-p}{p(1-\theta^*)}\Bigr)\Bigr\}\,\mathrm{d}\tau. \tag{34}$$ htl

We have $\widehat{\widetilde{\ell}}(\theta) = \widetilde{\ell}(\theta)$ for $\theta \le \theta^*$, and for $\theta \ge \theta^*$, $\widehat{\widetilde{\ell}}(\theta)$ is negative and decreasing (since $\theta^* > \theta_1$). Hence, $[\widehat{\widetilde{\ell}}(\theta)]^+ = \widetilde{\ell}^+(\theta)$ for all $0 < \theta \le 1$. In particular, $[\widehat{\widetilde{\ell}}(\theta)]^+ = \widetilde{\ell}^+(\theta) = 0$ for all $\theta \ge \theta^*$, and (33) implies

$$\max_{T^*<t\le n}\widehat{d}(t) = O_{L^2}(n^{1/2}). \tag{35}$$

Recalling $d(t) \le \widehat{d}(t)$, we thus have

$$\max_{T^*<t\le n}\bigl|d(t) - n\widetilde{\ell}^+(t/n)\bigr| = \max_{T^*<t\le n} d(t) \le \max_{T^*<t\le n}\widehat{d}(t) = O_{L^2}(n^{1/2}), \tag{36}$$

which completes the proof. □

Proof of Corollary 2.3. Immediate from Theorem 2.2 and (15), since we have $\max_t \widetilde{\ell}^+(t/n) = \max_\theta \widetilde{\ell}^+(\theta) + O(1/n)$ and $\max_\theta \widetilde{\ell}^+(\theta) = \widetilde{\ell}^+(\theta_0) = \widetilde{\ell}(\theta_0)$. □

Proof of Corollary 2.4. By Theorem 2.2,

$$\frac{1}{n}\sum_{t=1}^n d(t) = \sum_{t=1}^n \widetilde{\ell}^+(t/n) + O_{L^2}\bigl(n^{1/2}\bigr) = n\alpha + O_{L^2}\bigl(n^{1/2}\bigr), \tag{37}$$ bal1

where

$$\begin{aligned}\alpha &:= \int_0^1 \widetilde{\ell}^+(\tau)\,\mathrm{d}\tau = \int_0^{\theta_1}\widetilde{\ell}(\tau)\,\mathrm{d}\tau = \int_0^{\theta_1}\Bigl(\tau + \frac{1-p}{p}\log(1-\tau)\Bigr)\,\mathrm{d}\tau \\ &= \frac{1}{2}\theta_1^2 - \frac{1-p}{p}\Bigl((1-\theta_1)\log(1-\theta_1)+\theta_1\Bigr),\end{aligned} \tag{38}$$ bal2

which yields (22), using (16). □

2.4. **The trees in the forest.**

TT **Theorem 2.5.** *Let N be the number of trees in the depth-first forest. Then*

$$N = \rho n + O_{L^2}(n^{1/2}), \tag{39}$$ tt

where

$$\rho = \rho(p) := 1 - \theta_1 - \frac{p}{2(1-p)}(1-\theta_1)^2. \tag{40}$$ rho

Proof. Let $I_t := \mathbf{1}\{d(t) = 0\}$, the indicator that vertex t is a root and thus starts a new tree. Thus $N = \sum_1^n I_t$.

If $\theta_1 > 0$ (*i.e.*, $p > \frac{1}{2}$), then Theorem 2.2 shows that w.h.p. $d(t) > 0$ in the interval $(1, n\theta_1)$, except possibly close to the endpoints. Thus the DFS will find one giant tree of order $\approx \theta_1 n$, possibly preceded by a few small trees, and, as we will see later, followed by many small trees. To obtain a precise estimate, we note that there exists a constant $c > 0$ such that $\widetilde{\ell}(\theta) \ge \min\{c\theta, c(\theta_1 - \theta)\}$ for $\theta \in [0, \theta_1]$. Hence, if $t \le n\theta_1$ and $d(t) = 0$, then $\widetilde{d}(t) \le d(t) = 0$ by (8) and, recalling (25),

$$M^* \ge n\widetilde{\ell}(t/n) \ge c\min\{t, n\theta_1 - t\}. \tag{41}$$

Consequently, $d(t) = 0$ with $t \le n\theta_1$ implies $t \in [1, c^{-1}M^*] \cup [n\theta_1 - c^{-1}M^*, n\theta_1]$. The number of such t is thus $O(M^* + 1) = O_{L^2}(n^{1/2})$, using (25).

Let $T_1 := \lceil n\theta_1 \rceil$. We have just shown that (the case $\theta_1 = 0$ is trivial)

$$\sum_{t=1}^{T_1 - 1} I_t = O_{L^2}(n^{1/2}). \tag{42}$$ mma

It remains to consider $t \ge T_1$. Let

$$\mu_t := \mathbb{E}\,\xi_t = \frac{1 - \pi_t}{\pi_t} = \frac{1 - p}{p(1 - t/n)}. \tag{43}$$ mut

For any integer $k \ge 0$, the conditional distribution of $\xi_t - k$ given $\xi_t \ge k$ equals the distribution of ξ_t. Hence,

$$\mathbb{E}\big[(\xi_t - k)^+\big] = \mathbb{E}\big[\xi_t - k \mid \xi_t \ge k\big]\,\mathbb{P}(\xi_t \ge k) = \mu_t\,\mathbb{P}(\xi_t - k \ge 0). \tag{44}$$ erika

We use again the stochastic recursion (5). Let $\mathcal{F}_t$ be the σ-field generated by $\xi_1, \dots, \xi_{t-1}$. Then $d(t)$ is $\mathcal{F}_t$-measurable, while ξ_t is independent of $\mathcal{F}_t$. Hence, (5) and (44) yield

$$\begin{aligned}
\mathbb{E}\big[d(t+1) \mid \mathcal{F}_t\big] &= \mathbb{E}\big[d(t) + 1 - \xi_t \mid \mathcal{F}_t\big] + \mathbb{E}\big[(\xi_t - 1 - d(t))^+ \mid \mathcal{F}_t\big] \\
&= d(t) + 1 - \mu_t + \mu_t\,\mathbb{P}\big[\xi_t - 1 - d(t) \ge 0 \mid \mathcal{F}_t\big] \\
&= d(t) + 1 - \mu_t + \mu_t\,\mathbb{P}\big[d(t+1) = 0 \mid \mathcal{F}_t\big] \\
&= d(t) + 1 - \mu_t + \mu_t\,\mathbb{E}\big[I_{t+1} \mid \mathcal{F}_t\big].
\end{aligned} \tag{45}$$ ele

We write $\Delta d(t) := d(t+1) - d(t)$ and $\overline{I}_t := 1 - I_t$. Then (45) yields

$$\mathbb{E}\big[\Delta d(t) - 1 + \mu_t \overline{I}_{t+1} \mid \mathcal{F}_t\big] = 0. \tag{46}$$ win

Define

$$\mathcal{M}_t := \sum_{i=1}^{t-1} \mu_i^{-1}\big(\Delta d(i) - 1 + \mu_i \overline{I}_{i+1}\big) = \sum_{i=1}^{t-1}\Big(\mu_i^{-1}\Delta d(i) - \mu_i^{-1} + \overline{I}_{i+1}\Big). \tag{47}$$ cM

Then $\mathcal{M}_t$ is $\mathcal{F}_t$-measurable, and (46) shows that $\mathcal{M}_t$ is a martingale. We have, with $\Delta\mathcal{M}_t := \mathcal{M}_{t+1} - \mathcal{M}_t$, using (5),

$$|\Delta\mathcal{M}_t| \le \mu_t^{-1}\big|d(t+1) - d(t) - 1\big| + \overline{I}_{t+1} \le \mu_t^{-1}\xi_t + 1, \tag{48}$$ cm1

and thus, since $\pi_t \le p < 1$ for all t by (6),

$$\mathbb{E}|\Delta\mathcal{M}_t|^2 \le 2\mu_t^{-2}\,\mathbb{E}\,\xi_t^2 + 2 = 2\Big(\frac{\pi_t}{1 - \pi_t}\Big)^2 \frac{1 - \pi_t + (1 - \pi_t)^2}{\pi_t^2} + 2 = O(1). \tag{49}$$ cm2

Hence, uniformly for all $T \le n$,

$$\mathbb{E}\,\mathcal{M}_T^2 = \sum_{t=1}^{T-1} \mathbb{E}\,|\Delta\mathcal{M}_t|^2 = O(T) = O(n). \tag{50}$$ cm3

The definition (47) yields

$$\mathcal{M}_n - \mathcal{M}_{T_1} = \sum_{t=T_1}^{n-1} \mu_t^{-1}\Delta d(t) - \sum_{t=T_1}^{n-1} \mu_t^{-1} + \sum_{t=T_1}^{n-1} \overline{I}_{t+1}. \tag{51}$$ cm4

By a summation by parts, and interpreting $\mu_n^{-1} := 0$,

$$\sum_{t=T_1}^{n-1} \mu_t^{-1}\Delta d(t) = \sum_{t=T_1+1}^{n} \bigl(\mu_{t-1}^{-1} - \mu_t^{-1}\bigr) d(t) - \mu_{T_1}^{-1} d(T_1). \tag{52}$$ cm5

As t increases, μ_t increases by (43), and thus $\mu_{t-1}^{-1} - \mu_t^{-1} > 0$. Hence, (52) implies

$$\begin{aligned}\Bigl|\sum_{t=T_1}^{n-1} \mu_t^{-1}\Delta d(t)\Bigr| &\le \sum_{t=T_1+1}^{n} \bigl(\mu_{t-1}^{-1} - \mu_t^{-1}\bigr) \sup_{i>T_1} |d(t)| + \mu_{T_1}^{-1}|d(T_1)| \le 2\mu_{T_1}^{-1} \sup_{i\ge T_1} |d(t)| \\ &= O_{L^2}(n^{1/2}) \end{aligned} \tag{53}$$ cm6

by (18), since $\widetilde{\ell}^{+}(t/n) = 0$ for $t \ge T_1 \ge n\theta_1$. Furthermore, (50) shows that $\mathcal{M}_n, \mathcal{M}_{T_1} = O_{L^2}(n^{1/2})$. Hence, (51) yields

$$\sum_{t=T_1+1}^{n} I_t = n - T_1 - \sum_{t=T_1+1}^{n} \overline{I}_t = n - T_1 - \sum_{t=T_1}^{n-1} \mu_t^{-1} + O_{L^2}(n^{1/2}) = n\rho + O_{L^2}(n^{1/2}), \tag{54}$$ cm7

where, recalling (43),

$$\rho := 1 - \theta_1 - \int_{\theta_1}^{1} \frac{p(1-\tau)}{1-p}\,\mathrm{d}\tau = 1 - \theta_1 - \frac{p}{2(1-p)}(1-\theta_1)^2. \tag{55}$$ rho=

The result follows by (54) and (42). □

The argument in the proof of Theorem 2.5 shows also the following; we omit the details.

TT2 **Theorem 2.6.** *If $p > \frac{1}{2}$, then the largest tree in the depth-first forest has order $\theta_1 n + O_{L^2}(n^{1/2})$.*

Rslow **Remark 2.7.** When $p > \frac{1}{2}$, the height is thus linear in n, unlike many other types of random trees. This might imply a rather slow performance of algorithms that operate on the depth-first forest.

Conjecture 2.8. *If $p = \frac{1}{2}$, then the largest tree has order roughly $n^{2/3}$. If $p < \frac{1}{2}$, then the largest tree has order roughly* $\log n$.

2.5. **Types of arcs.** Recall from the introduction the classification of the arcs in the digraph G. Since we assume that the outdegrees are $\mathrm{Ge}(1-p)$ and independent, the total number of arcs, M say, has a negative binomial distribution with mean λn, and, by a weak version of the law of large numbers,

$$M = \lambda n + O_{L^2}(n^{1/2}). \tag{56}$$ sw

In the following theorem, we give the asymptotics of the expected number of arcs of each type. We conjecture that $\operatorname{Var} B, \operatorname{Var} F, \operatorname{Var} C = O(n)$, so that (59)–(60) can be improved to asymptotics for the random numbers similar to (58).

Tarcs **Theorem 2.9.** *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 G.*

$$L = O_{L^2}(1), \tag{57}$$ tal

$$T = \chi n + O_{L^2}(n^{1/2}), \tag{58}$$ tat

$$\mathbb{E}\, F = \mathbb{E}\, B = \lambda\, \mathbb{E}\, \overline{d} = \beta n + O(n^{1/2}), \tag{59}$$ taf

$$\mathbb{E}\, C = \chi n + O(n^{1/2}), \tag{60}$$ tac

where

$$\beta := \lambda\alpha = \frac{2p-1}{1-p}\theta_1 - \frac{\lambda}{2}\theta_1^2, \tag{61}$$ gb

$$\chi := 1-\rho = \theta_1 + \frac{\lambda}{2}(1-\theta_1)^2. \tag{62}$$ chi

Remark 2.10. Part of the theorem is the exact equality $\mathbb{E}\,B = \mathbb{E}\,F$ for any n and p. Knuth [1] conjectures, based on exact formulas for small n, that, much more strongly, B and F have the same distribution for every n.

Moreover, Knuth [1] conjectures that $\mathbb{E}\,C = \mathbb{E}\,T$ for every n. (The theorem above shows only that this holds asymptotically as $n \to \infty$.)

Proof. L: A simple argument with generating functions shows that the number of loops at a given vertex v is $\mathrm{Ge}(1-p/(n-np+p))$; these numbers are independent, and thus $L \sim \mathrm{NegBin}\bigl(n, 1-p/(n-np+p)\bigr)$ with $\mathbb{E}\,L = p/(1-p) = O(1)$ and $\mathrm{Var}(L) = p(1-p+p/n)/(1-p)^2 = O(1)$ [1]. Moreover, it is easily seen that asymptotically, L has a Poisson distribution, $L \xrightarrow{\mathrm{d}} \mathrm{Po}(\lambda)$

T: This follows immediately from Theorem 2.5, since $T = n - N$.

B, F: Let v, w be two distinct vertices. If the DFS finds w as a descendant of v, then there will later be $\mathrm{Ge}(1-p)$ arcs from w, and each has probability $1/n$ of being a back arc to v. Similarly, there will be $\mathrm{Ge}(1-p)$ future arcs from v, and each has probability $1/n$ of being a forward arc to w. Hence, if I_{vw} is the indicator that w is a descendant of v, and B_{vw} $[F_{vw}]$ is the number of back [forward] arcs vw, then

$$\mathbb{E}\,B_{wv} = \mathbb{E}\,F_{vw} = \frac{\lambda}{n}\,\mathbb{E}\,I_{vw}. \tag{63}$$

Summing over all pairs of distinct v and w, we obtain

$$\mathbb{E}\,B = \mathbb{E}\,F = \frac{\lambda}{n}\,\mathbb{E}\sum_w \sum_{v\neq w} I_{vw} = \frac{\lambda}{n}\,\mathbb{E}\sum_w d(w), \tag{64}$$

and (59) follows by Corollary 2.4.

C: We have $L+T+B+F+C = M$, and thus (60) follows from (56) and (57)–(59), noting that (61)–(62) imply $\beta+\chi = \lambda/2$, and thus $\lambda - (\chi+2\beta) = \chi$. □

3. Depth, trees and arc analysis in the shifted geometric outdegree distribution

In this section, the outdegree distribution is $\mathrm{Ge}_1(1-p) = 1 + \mathrm{Ge}(1-p)$. Thus its mean $\lambda = 1/(1-p)$. As in Section 2, the depth $d(t)$ is a Markov chain given by (5), but the distribution of ξ_t is now different. The probability (1) is replaced by $(1-t/n)/(1-pt/n)$, but the number of future arcs from an ancestor is still $\mathrm{Ge}(1-p)$, and, with $\theta := t/n$,

$$\mathbb{P}\bigl(\xi_t = k\bigr) = \begin{cases} \overline{\pi}_t := \frac{1-\theta}{1-p\theta}, & k = 0, \\ (1-\overline{\pi}_t)(1-\pi_t)^{k-1}\pi_t, & k \geq 1, \end{cases} \tag{65}$$ xi1

where $\pi_t = p\overline{\pi}_t$ is as in (6). The probability generating function of ξ_t is, instead of (9),

$$p(t,z) = \overline{\pi}_t + (1-\overline{\pi}_t)\frac{\pi_t z}{1-(1-\pi_t)z} = (1-\theta)\frac{1-(1-p)z}{1-p\theta-(1-p)z}. \tag{66}$$

We now have $\mathbb{E}\,\xi_i = \frac{(1-p)\theta}{p(1-\theta)}$ and instead of (14) we have $\mathbb{E}\,\widetilde{d}(t) = n\widetilde{\ell}(\theta) + O(1)$ where now $\widetilde{\ell}(\theta)$ takes the new value

$$\widetilde{\ell}(\theta) := \frac{1}{p}\theta + \frac{1-p}{p}\log(1-\theta) \tag{67}$$ tl1

The rest of the analysis does not change but we get different values for the constants. Note that $\widetilde{\ell}(\theta_1) = 0$ still gives (16), now with $\lambda = 1/(1-p)$, and that $\lambda > 1$ for every p. Differentiating (67) shows that the maximum point $\theta_0 = p > 0$.

The results in Theorems 2.2–2.6 thus hold, with, by straightforward calculations,

$$\upsilon := \widetilde{\ell}(p) = 1 + \frac{1-p}{p}\log(1-p), \tag{68}$$

$$\alpha := \frac{1}{p}\left(\frac{\theta_1^2}{2} - \frac{1}{\lambda}((1-\theta_1)\log(1-\theta_1) - \frac{1}{\lambda}\theta_1\right) = \theta_1 - \frac{\theta_1^2}{2p}, \tag{69}$$

$$\rho := 1 - \theta_1 - \frac{1}{2(1-p)}(1-\theta_1)^2. \tag{70}$$

Figure 2 shows $\widetilde{\ell}(\theta)$ for both geometric distributions.

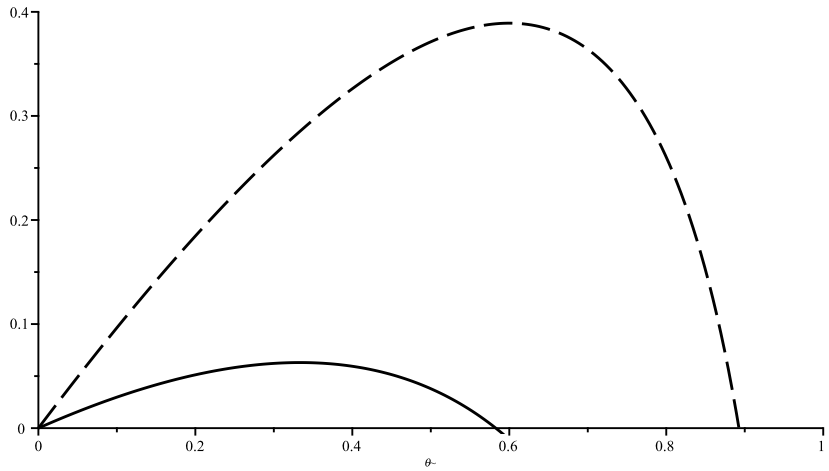

FIGURE 2. $\widetilde{\ell}(\theta)$, the asymptotic search depth, for geometric distribution (solid) and shifted geometric distribution (dashed) with $p = 0.6$.

fig:depth

Now the expected numbers of back and forward arcs differ since $\mathbb{E}\,B = \lambda\,\mathbb{E}\,\overline{d} \sim \lambda\alpha n$ and $\mathbb{E}\,F = (\lambda-1)\,\mathbb{E}\,\overline{d} \sim (\lambda-1)\alpha n$ because the average number of future arcs at a vertex after a descendant have been created is $\lambda - 1$. Thus the equality of the number of backward and forward arcs in Theorem 2.9 was an artefact of the geometric degree distribution.

The number of tree arcs is still

$$T = \chi n + O_{L^2}(n^{1/2}) \tag{71}$$

with $\chi = 1 - \rho = \theta_1 + \frac{\lambda}{2}(1-\theta_1)^2$. But the number of cross arcs differs. We get

$$\mathbb{E}\,C = \chi' n + O(n^{1/2}) \tag{72}$$

with

$$\chi' := \lambda - \chi - \lambda\alpha - (\lambda-1)\alpha = \frac{\lambda}{2}\Big(1 - 2\theta_1 + \frac{1}{p}\theta_1^2\Big) = \frac{\lambda}{2}(1-\theta_1)^2 + \frac{1}{2p}\theta_1^2. \tag{73}$$

Again the (asymptotic) identity of $\mathbb{E}\,T$ and $\mathbb{E}\,C$ was an effect of the geometric distribution.

4. Stack index analysis and forest size for a general outdegree distribution

In this section, we consider a general outdegree distribution, with mean λ and finite variance. When the outdegree distribution is general, the depth does not longer follow an easy Markov chain, since we should keep track of the number of children seen so far at each level of the branch of the tree toward the current vertex.

We denote m the degree of a random vertex. When the distribution of m is general, the depth does not longer follow an easy Markov chain, since we should keep track of the number of children seen so far at each level of the branch of the tree toward the current vertex. However we can set the following theorem.

Theorem 4.1. *In the general outdegree distribution when $\lambda > 1$ we have the following expansion for small values of t:*

$$\mathbb{E}[d(t)] = (1 - T(1))t + O(t^2/n) \tag{74}$$

with $T(z)$ the p.g.f. of the finite tree of the Galton Watson tree generated by the outdegree m.

Remark 4.2. It turns out that the theorem also holds for $\lambda < 1$, but in this case $T(1) > 1$.

Proof. The p.g.f. $T(z)$ satisfies $T(z) = z\,\mathbb{E}[T(z)^m]$. For the geometric distribution, when $p > 0.5$ we have

$$T(z) = \frac{1 - \sqrt{1 - 4(1-p)pz}}{2p} \tag{75}$$

and $T(1) = \frac{1-p}{p}$ is the probability to have a finite tree. For the shifted geometric distribution we simply have $T(z) = 0$. If we omit the variation of the fraction of discovered vertices θ, the tree search experience starts by a sequence of finite trees, before entering into the infinite tree. Said in other words, the depth reaches the level 1 and will never come back to zero. But it will hit level 1 a finite time before leaving it for ever. If the degree of the root is m, the search is limited to m subtrees among them there are one or more infinite trees. The probability that the first infinite sub-tree is the kth sub-tree is $T(1)^{k-1}(1 - T(1))$, therefore the cumulated size of the finite trees before the first infinite subtree is $\frac{1-T(z)^m}{1-T(z)}(1 - T(1))$, multiplying by z we get the p.g.f. of the duration to reach the ultimate level 2 (*i.e.* the level which will never be decremented afterward). Averaging over the distribution of degrees and normalizing on the fact that there is always an infinite sub-tree we get $\frac{z - z\,\mathbb{E}[T(z)^m]}{1-T(z)}$. Since $z\,\mathbb{E}[T(z)^m] = T(z)$ we get $\frac{z-T(z)}{1-T(z)}$ whose average obtained by the first derivative at $z = 1$ is exactly $\frac{1}{1-T(1)}$.

The average duration from the ultimate level 2 to the ultimate is also $\frac{1}{1-T(1)}$. And so forth to reach the ultimate level k we need an average duration $\frac{k}{1-T(1)}$. Therefore we have

$$\mathbb{E}[d(t)] = (1 - T(1))t(1 + O(\theta)) \tag{76}$$

the term $O(\theta)$ comes from the impact of the variation of θ when t is small which have neglected in the first place. □

However we don't know how to go further. If we want to get back to a Markov chain we must replace the depth by the stack index $I(t)$. The DFS can be regarded as keeping a stack of unexplored arcs, for which we have seen the start vertex but not the end. The stack evolves as follows:

`stack1`

(1) 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.
(2) (v is now a new vertex) Reveal the outdegree m of v and add to the stack m new arcs, with unspecified endpoints. GOTO 1.

Let again v_t be the tth vertex seen by the DFS, and let $I(t)$ be the size of the stack when $v(t)$ is found (but before we add the arcs from v_t). Also let η_t be the outdegree of v_t. Then $I(1) = 0$ and, in analogy with (5),

$$I(t+1) = \bigl(I(t) + \eta_t - 1 - \xi_t\bigr)^+, \qquad 1 \le t < n, \tag{77}$$ It

where ξ_t is the number of arcs leading to already seen vertices before we find a new one; we have $\mathbb{P}(\xi = k) = (1 - \frac{t}{n})(\frac{t}{n})^k$ and thus $\xi_t \sim \mathrm{Ge}(1 - t/n)$.

In analogy with (7), we define also

$$\widetilde{I}(t) := \sum_{i=1}^{t-1} (\eta_i - 1 - \xi_i) = \sum_{i=1}^{t-1} \zeta_i, \tag{78}$$ tIt

where we define $\zeta_t := \eta_t - \xi_t - 1$. Then, as in (8),

$$I(t) = \widetilde{I}(t) - \min_{1 \le j \le t} \widetilde{I}(j). \tag{79}$$ ItI

Note that

$$\mathbb{E}\,\zeta_t = \mathbb{E}\,\eta_t - \mathbb{E}\,\xi_t - 1 = \lambda - \frac{t/n}{1 - t/n} - 1 = \lambda - \frac{1}{1 - t/n}. \tag{80}$$

Hence, uniformly in $t/n \le \theta^*$ for any fixed $\theta^* < 1$,

$$\mathbb{E}\,\widetilde{I}(t) = \sum_{i=1}^{t-1} \mathbb{E}\,\zeta_i = (t-1)\lambda - \sum_{i=1}^{t-1} \frac{1}{1 - t/n} = n\widetilde{\iota}(t/n) + O(1), \tag{81}$$

where

$$\widetilde{\iota}(\theta) := \int_0^\theta \Bigl(\lambda - \frac{1}{1-\tau}\Bigr)\,\mathrm{d}\tau = \lambda\theta + \log(1-\theta). \tag{82}$$

Let

$$\widetilde{\iota}^+(\theta) := [\widetilde{\iota}(\theta)]^+ = \begin{cases} \lambda\theta + \log(1-\theta), & 0 \le \theta \le \theta_1, \\ 0, & \theta_1 \le \theta \le 1, \end{cases} \tag{83}$$

where again θ_1 is the largest root in $[0,1)$ of (16), now with $\lambda = \mathbb{E}\,\eta_1$, the mean of $\mathbf{P}$. The proof of Theorem 2.2 applies with very minor differences, and yields:

TG1 **Theorem 4.3.** *Suppose that the outdegree distribution has finite variance. Then*

$$\max_{1 \le t \le n} \bigl|I(t) - n\widetilde{\iota}^+(t/n)\bigr| = O_{L^2}(n^{1/2}). \tag{84}$$

Moreover, v_{t+1} is a root if and only if $I(t) + \zeta_t = I(t) + \eta_1 - 1 - \xi_t < 0$, cf. (77). The arguments in the proof of Theorem 2.5 apply with minor differences, and show:

Theorem 4.4. *Theorems 2.5 and 2.6 hold for any outdegree distribution with finite variance, with* $\rho := 1 - \theta_1 - \frac{\lambda}{2}(1-\theta_1)^2$.

The previous results suggest that the $\mathbb{E}\,\widetilde{I}(t)$ and $\mathbb{E}\,\widetilde{d}(t)$ asymptotically are just proportional by a fixed factor independent of t; note that both have the same root θ_1. We make the following conjecture.

Conjecture 4.5. *If $\lambda > 1$, then, for all $t \in [0,1]$, we have $\mathbb{E}\,\widetilde{d}(t) = \frac{1-z_1}{\lambda-1}\,\mathbb{E}\,\widetilde{I}(t) + o(n)$ where $z_1 \in [0,1)$ satisfies $z_1 = \mathbb{E}[z_1^\eta]$.*

Remark 4.6. The quantity z_1 is the probability that a Galton–Watson tree generated by the outdegree distribution η is finite. The reason to invoke this probability is that $\mathbb{E}\,\widetilde{d}(t)$ should be asymptotically equivalent to θ times the probability of an infinite tree $(1 - z_1)$ when $\theta \to 0$. For the geometric distribution the probability is $\frac{1-p}{p}$ thus with $\lambda = \frac{p}{1-p}$ the proportional factor is $\frac{1-p}{p}$ and we find again our first result. When $\lambda < 1$ the fixed point

$z_1 > 1$ and the proportional factor remains positive. For the shifted geometric distribution the Galton–Watson tree is always in finite, thus $z_1 = 0$ and the proportional factor is again $\frac{1-p}{p}$.

References

Knuth12A [1] Donald E. Knuth, *The Art of Computer Programming*, Section 7.4.1.2 (Preliminary draft, 13 December 2021). http://cs.stanford.edu/~knuth/fasc12a.ps.gz

Inria Saclay Ile de France, France
Email address: philippe.jacquet@inria.fr

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
Email address: svante.janson@math.uu.se
URL: http://www2.math.uu.se/~svante/