ON KNUTH'S CONJECTURE FOR BACK AND FORWARD ARCS IN DEPTH-FIRST SEARCH IN A RANDOM DIGRAPH WITH GEOMETRIC OUTDEGREE DISTRIBUTION

SVANTE JANSON

To Donald Knuth on his 85th birthday

ABSTRACT. Donald Knuth, in a draft of a coming volume of *The Art of Computer Programming*, has recently conjectured that in Depth-First Search of a random digraph with geometric outdegree distribution, the numbers of back and forward arcs have the same distribution.

We show that this conjecture is equivalent to an equality between two generating functions defined by different recursions.

Unfortunately, we have not been able so use this to prove the conjecture, which still is open, but we hope that this note will inspire others to succeed with the conjecture.

1. INTRODUCTION

Donald Knuth [2, Section 7.4.1.2], in a draft of a coming volume of *The Art of Computer Programming*, studies Depth-First Search (DFS) in a digraph. Among the results there, he makes an intriguing conjecture [2, Problem 7.4.1.2-35] for the case when the digraph is random with a geometric outdegree distribution; see Section 2 below. The conjecture is strongly supported by exact calculations showing that it holds for small digraphs (order $n \leqslant 9$, extend by us by the methods below to $n \leqslant 18$).

We have not been able to prove this conjecture, but we show below (Corollary 3.3) that the conjecture is equivalent to the equality of two generating functions defined by similar but different recursions. (We also give an extended version of the conjecture.) The purpose of this note is to present this reformulation in the hope of stimulating others to show the conjecture.

2. KNUTH'S CONJECTURE

2.1. **Depth-First Search.** For a detailed description of DFS in a digraph, and for historical notes, see [2, Section 7.4.1.2]. Briefly, the DFS starts with an arbitrary vertex, and explores the arcs from that vertex one by one. When an arc is found leading to a vertex that has not been seen before, the DFS explores the arcs from it in the same way, in a recursive fashion, before returning to the next arc from its parent. This eventually yields a tree containing all descendants of the the first vertex. If there still are some unseen vertices, the DFS starts again with one of them and finds a new tree, and so on until all vertices are found. The DFS thus yields a spanning forest in the digraph, called the *depth-first forest*.

The discarded arcs (leading to already seen vertices) may be classified further; [2, Section 7.4.1.2] classifies the arcs in the digraph into the following five types

- *loops*;
- *tree arcs*, the arcs in the resulting depth-first forest;
- *back arcs*, the arcs that point to an ancestor of the current vertex in the current tree;
- *forward arcs*, the arcs that 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).

Each instance of DFS in a digraph is described by the sequence of arcs that are explored, taken in the order they are explored. For an instance of the DFS, let L, F, B, C, T be the numbers of loops, forward, backward, cross, and tree arcs.

Let $F_n(w, x, y, z, t)$ be the generating function defined as the sum of $w^L x^F y^B z^C t^T$ over all instances of a DFS of a digraph on n vertices labelled $1, \ldots, n$, where we for definiteness assume that when the DFS starts a new tree, it chooses as the root the unused vertex with smallest label. (Algorithm D in [2, 7.4.1.2] chooses the unused vertex with largest label; this is obviously equivalent for our purposes, but notationally less convenient for us.)

2.2. **Random digraphs with geometric outdegree distribution.** From now on we consider DFS in a random digraph given by the following model.

The number of vertices n is given; we are also given a parameter $p \in (0, 1)$. Each vertex is given an outdegree that is a random variable with a geometric distribution $\mathrm{Ge}(1 - p)$; in other words, if the outdegree of vertex i is η_i, then

$$\mathbb{P}(\eta_i = k) = p^k(1 - p), \qquad k = 0, 1, 2, \ldots; \tag{2.1}$$

moreover, the outdegrees of different vertices are independent. This gives each vertex i a number η_i of outgoing arcs; the other endpoint of these arcs are chosen uniformly at random among all n vertices, independently for all arcs. (Thus, loops and multiple arcs are allowed, so the digraph is really a multidigraph.)

2.3. **The conjecture.** Donald Knuth [2, Problem 7.4.1.2-35] conjectures that for a digraph with a geometric outdegree distribution, for every $n \geqslant 1$ and every $p \in (0, 1)$, the random variables F and B have the same distribution, which we write as

$$F \overset{\mathrm{d}}{=} B. \tag{2.2}$$

Equivalently, in terms of probability generating functions,

$$\mathbb{E} \, x^F = \mathbb{E} \, x^B. \tag{2.3}$$

(Knuth verified this for $n \leqslant 9$ by explicit calculations.)

2.4. **The conjecture and generating functions.** Given a sequence of arcs that describes an instance of a DFS, we find the probability of observing exactly this sequence as follows: for each arc in the sequence, the emitting vertex should send out at least one more arc after what we may have seen earlier (probability p), and the endpoint of this arc should be a specified vertex (probability $1/n$); furthermore, for each vertex, there is one time when this vertex had a chance to send out more vertices but did not do so (probability $1 - p$). Hence, the probability is $(p/n)^{L+F+B+C+T}(1-$

$p)^n$. Consequently, the joint probability generating function of (L, F, B, C, T) for the DFS in a random digraph with n vertices and $\eta \sim \mathrm{Ge}(1-p)$ is

$$\mathbb{E}\big[w^L x^F y^B z^C t^T\big] = (1-p)^n F_n\Big(w\frac{p}{n}, x\frac{p}{n}, y\frac{p}{n}, z\frac{p}{n}, t\frac{p}{n}\Big). \tag{2.4}$$

By (2.4), Knuth's conjecture (2.2) is thus equivalent to

$$F_n\Big(\frac{p}{n}, x\frac{p}{n}, \frac{p}{n}, \frac{p}{n}, \frac{p}{n}\Big) = F_n\Big(\frac{p}{n}, \frac{p}{n}, x\frac{p}{n}, \frac{p}{n}, \frac{p}{n}\Big), \tag{2.5}$$

or, by first replacing p/n by w and then xw by x,

$$F_n\big(w, x, w, w, w\big) = F_n\big(w, w, x, w, w\big). \tag{2.6}$$

In fact, we conjecture, more generally, that for every $n \geqslant 1$,

$$F_n\big(w, x, z, z, t\big) = F_n\big(w, z, x, z, t\big), \tag{2.7}$$

which by the same argument is equivalent to

$$\mathbb{E}\big[w^L x^F z^{B+C} t^F\big] = \mathbb{E}\big[w^L x^B z^{F+C} t^F\big], \tag{2.8}$$

and thus to the following extended version of Knuth's conjecture, with joint distributions of different variables.

Conjecture 2.1. *For every $n \geqslant 1$ and every $p \in (0, 1)$,*

$$(L, F, B + C, T) \overset{\mathrm{d}}{=} (L, B, F + C, T). \tag{2.9}$$

Remark 2.2. The much weaker statement that $\mathbb{E}\, F = \mathbb{E}\, B$, i.e., that the expectations are the same, is proved in [1, Theorem 2.18]. It is there also shown that $\mathbb{E}\, T = \mathbb{E}\, C$, but that T and C do not have the same distribution. $\qquad\square$

Remark 2.3. It is easy to see that (2.9) cannot be extended to $(L, F, B, C, T) \overset{\mathrm{d}}{=} (L, B, F, C, T)$. This fails already for $n = 3$, as can be seen by the argument in Section 3, calculating $G_3(w, x, y, z)$ by (3.6) and noting that $G_3(w, x, y, z) \neq G_3(w, y, x, z)$. $\qquad\square$

3. Recursion formulas

To find F_n, consider now the DFS stopped when the first tree is completed, and define $G_m(w, x, y, z, t)$ as the sum of $w^L x^F y^B z^C t^T$ over all sequences of arcs that yield the first tree in an instance of DFS, where moreover this tree has m given vertices and these are visited in a given order (for example, the vertices $1, 2, \ldots, m$); we here define L, F, B, C, T as above also for sequences of arcs that are only a part of the full DFS.

In general, a DFS creates a forest with several trees. Consider the contribution to F_n from the case when the depth-first forest have k given trees with sizes $m_1, m_2, \ldots, m_k$, with vertices visited in a given order. The exploration of the first tree gives a factor $G_{m_1}(w, x, y, z, t)$. The exploration of the second tree proceeds like an exploration of just that tree, but we may also have cross arcs that go back to the any of the m_1 vertices of the first tree; in fact, any time that we may have a loop we may instead have one of these cross arcs, and therefore the exploration of the second tree yields a factor $G_{m_2}(w + m_1 z, x, y, z, t)$, where w is replaced by $w + m_1 z$ to account for the possible additional cross arcs. The same reasoning applies to all following trees, and it follows that the contribution to $F_n(w, x, y, z, t)$ is $\prod_{i=1}^{k} G_{m_i}(w + M_{i-1} z, x, y, z, t)$, where $M_j := \sum_{i \leqslant j} m_i$, and thus $F_n(w, x, y, z, t)$

equals the sum over all possible depth-first forests of such products. Consequently, (2.7) follows if, for every $m \geqslant 1$,

$$G_m\big(w, x, z, z, t\big) = G_m\big(w, z, x, z, t\big). \tag{3.1}$$

Conversely, it follows by induction that (3.1) is necessary for (2.7) to hold for all n.

In the definition of G_m, we consider only sequences of arcs that produce a tree with m vertices; thus the number of tree edges $T = m - 1$. Hence we have

$$G_m(w, x, y, z, t) = t^{m-1} G_m(w, x, y, z, 1). \tag{3.2}$$

We simplify the notation by writing $G_m(w, x, y, z) := G_m(w, x, y, z, 1)$, and see that (3.1) is equivalent to

$$G_m\big(w, x, z, z\big) = G_m\big(w, z, x, z\big). \tag{3.3}$$

We find a recursion formula for G_n (now replacing m by n for convenience). For $n = 1$, the sequences of arcs that appear in the definition of G_1 are the sequences of 0 or more loops at the root. Hence, $F = B = C = 0$ and

$$G_1(w, x, y, z) = \sum_{i=0}^{\infty} w^i = \frac{1}{1 - w}. \tag{3.4}$$

For $n \geqslant 2$, suppose that the DFS produces a tree τ on the vertices $1, \ldots, n$ (in this order). Consider the last child of the root, and let it have label $m + 1$ where $1 \leqslant m \leqslant n - 1$. Let τ' be the subtree of τ formed by the vertices before $m + 1$, i.e. $1, \ldots, m$, and let τ'' be the subtree formed by $m + 1$ and its descendants; these have sizes $|\tau'| = m$ and $|\tau''| = n - m$. An exploration of τ consists of

(1) an exploration of τ',
(2) a tree arc from the root 1 to $m + 1$,
(3) an exploration of τ'',
(4) return to the root and a number of arcs leading only to already seen vertices.

In order for a sequence of arcs to produce the tree τ, we have as long as we are exploring inside τ' exactly the same possibilities as for the tree τ' on its own, but during the exploration of τ'', we can also have edges to vertices in τ'; more precisely, we see that for any exploration yielding the tree τ'', each time we may have a loop, we can in the exploration of τ also have a back edge to the root 1, or a cross edge to one of the $m - 1$ other vertices in τ'. This corresponds to replacing w by $w + y + (m - 1)z$ in the generating function. Moreover, in step 4, we can have any number $\geqslant 0$ of arcs from 1 to itself (a loop) or to one of the $n - 1$ other vertices in τ (a forward arc). This yields for the generating function a factor

$$\sum_{i=0}^{\infty} (w + (n-1)x)^i = \frac{1}{1 - w - (n-1)x}. \tag{3.5}$$

Consequently, summing over all possible m and trees τ' and τ'', we find, for $n \geqslant 2$,

$$G_n(w, x, y, z) = \frac{1}{1 - w - (n-1)x} \sum_{m=1}^{n-1} G_m(w, x, y, z) G_{n-m}(w + y + (m-1)z, x, y, z). \tag{3.6}$$

We have shown the following:

Proposition 3.1. *The generating function G_n satisfies the recursion* (3.6) *with the initial value* (3.4).

Specializing to the parameters in (3.3), we have thus the following corollary. Here

$$\widehat{G}_n(w,x,z) := G_n(w,x,z,z) \quad \text{and} \quad \widecheck{G}_n(w,x,z) := G_n(w,z,x,z) \qquad (3.7)$$

are generating functions for $(L, F, B + C)$ and $(L, B, F + C)$, respectively.

Proposition 3.2. *Let $\widehat{G}_n(w,x,z)$ and $\widecheck{G}_n(w,x,z)$ be defined by the recursions*

$$\widehat{G}_1(w,x,z) := \widecheck{G}_1(w,x,z) := \frac{1}{1-w} \qquad (3.8)$$

and, for $n \geqslant 2$,

$$\widehat{G}_n(w,x,z) := \frac{1}{1-w-(n-1)x} \sum_{m=1}^{n-1} \widehat{G}_m(w,x,z)\widehat{G}_{n-m}(w+mz,x,z), \qquad (3.9)$$

$$\widecheck{G}_n(w,x,z) := \frac{1}{1-w-(n-1)z} \sum_{m=1}^{n-1} \widecheck{G}_m(w,x,z)\widecheck{G}_{n-m}(w+x+(m-1)z,x,z). \qquad (3.10)$$

Then (2.9), or equivalently (2.7), holds if and only if

$$\widehat{G}_n(w,x,z) = \widecheck{G}_n(w,x,z), \qquad n \geqslant 1. \qquad (3.11)$$

Proof. First, $\widehat{G}_n$ and $\widecheck{G}_n$ satisfy (3.8)–(3.10) by (3.7) and (3.6), so they can be defined by these recursions.

Furthermore, we have shown that (2.9) $\iff$ (2.7) $\iff$ (3.1) $\iff$ (3.3), and (3.3) $\iff$ (3.10) by (3.7). $\qquad\square$

In order to show (2.6), i.e. Knuth's conjecture, it suffices to verify (3.11) for $z = w$. In other words:

Corollary 3.3. *Knuth's conjecture $F \overset{\mathrm{d}}{=} B$, or equivalently (2.6), is equivalent to*

$$\widehat{G}_n(w,x,w) = \widecheck{G}_n(w,x,w), \qquad n \geqslant 1, \qquad (3.12)$$

where $\widehat{G}_n$ and $\widecheck{G}_n$ are given by the recursions (3.8)–(3.10).

Problem 3.4. Prove (3.11), or at least (3.12)!

Remark 3.5. Knuth verified his conjecture by exact calculations for $n \leqslant 9$. We have extended this by verifying (3.11) (and thus (2.7)) for $n \leqslant 18$. $\qquad\square$

Remark 3.6. Another recursion for $\widecheck{G}(w,x,z)$ is

$$\widecheck{G}_n(w,x,z) := \frac{1}{1-w} \sum_{m=1}^{n-1} \widecheck{G}_m(w+x,x,z)\widecheck{G}_{n-m}(w+mz,x,z). \qquad (3.13)$$

This is obtained similarly as the recursion above, but now defining τ' to be the subtree of τ consisting of the first child of the root (i.e. 2) and all its descendants, and τ'' to be the subtree consisting of all other vertices (including the root 1). An exploration of τ consists of

(1) a number $\geqslant 0$ of loops at the root 1,
(2) a tree arc from 1 to 2,
(3) an exploration of τ',
(4) return to the root and an exploration of τ''.

6 SVANTE JANSON

Step (1) yields a factor $(1 - w)^{-1}$ in the generating function. Step (3) is as an exploration of the tree τ', but each time we may have a loop, we may also have a back edge to 1; this corresponds to replacing w by $w + y$ in the generating function. Similarly, in Step (4), each time we may have a loop, we may also have an edge to one of the vertices in τ'; this is either a forward edge (if we are at the root) or a cross edge; this corresponds to replacing w by $w + mz$ in the generating function, with $m := |\tau'|$. Together, this yields (3.13).

Note that although it seems possible to derive also a recursion for $\widehat{G}_n(w, x, z)$, or more generally for $G_n(w, x, y, z)$, by the decomposition used here, it will be more complicated because we then would have to keep control over whether edges from τ'' to τ' are forward or cross edges. We leave this to the reader. $\qquad\square$

References

[1] Philippe Jacquet & Svante Janson. Depth-First Search performance in a random digraph with geometric outdegree distribution. Preprint, 2022. `arXiv:2212.14865`

[2] Donald E. Knuth. *The Art of Computer Programming*, Section 7.4.1.2 (Preliminary draft, 13 February 2022). `http://cs.stanford.edu/~knuth/fasc12a.ps.gz`

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/`