NOTES ON GROWING A TREE IN A GRAPH

February 9, 2017

1 Introduction

We consider the following process for growing a spanning tree, T, of a graph G starting at some vertex $s \in V(G)$. Intially, $T = (s, \emptyset)$ is the single vertex tree containing only s. We then repeatedly select, uniformly at random, an edge from $E(G)$ that has one endpoint in $V(T)$ and one endpoint not in $V(T)$ and we add this edge to T. For an n-vertex connected graph G, the tree T spans G after $n-1$ steps. We call this Process A. We are interested in the height of the (random) spanning tree generated by Process A.

It turns out that there are several equivalent views of Process A. Consider the following, which we call Process E (for exponential). On each edge of G we attach an exponential(1) timer. When the timer on an edge vw rings the timer is immediately reset and, if exactly one of v or w is in T, then the edge vw is added to T. We say that Process E is *complete* once T spans G. Note that, by the memorylessness of exponential random variables, at any point in time, each edge is equally likely to be the next edge whose timer rings. Thus, Process E produces spanning trees with the same distribution as those produced by Process A.

Also, by the memorylessness of exponential random variables, Process E is equivalent to selecting an exponential(1) edge weight for each edge of G and then computing the shortest path tree rooted at s. We call this latter process *Process FP* (for first-passage percolation).

Since these processes produce the same distribution of spanning trees, in the remainder, T will refer to a spanning tree produced by Process A, Process E, or Process FP, whichever is convenient. Since our Process A refers to an unweighted graph and Process FP refers to weighted graph, we will use the convention that the *length* of a path P is the number of edges in the path and the *weight*, $W(P)$ of a path is the sum of the weights on the edges in the path. The *height*, $h(T)$, of T is the length of the longest root-to-leaf path in T.

In this note, we show that the height of the tree T is (with high probability) $O(\Delta(D+\log n))$, where n is the number of vertices in G, D is the diameter of G, and Δ is the maximum degree of any vertex in G. We also show that this bound is essentially tight by presenting a family of graphs for which the the height of T is (with high probability) $\Omega(\Delta D)$.

2 Inequalities for Sums of Exponentials

We will make use of two inequalities on sums of exponential random variables, both of which can be obtained using Chernoff's bounding method. If $Z_1,\dots,Z_k$ are independent exponential(λ) random variables (so that they each have mean $\mu = 1/\lambda$), then for all $d > 1$,

$$\Pr\left\{\sum_{i=1}^{k} Z_i \le \mu k/d\right\} \le \exp(-k(\ln d - 1 + 1/d)) \le \left(\frac{e}{d}\right)^k \tag{1}$$

and for all $t > 1$,

$$\Pr\left\{\sum_{i=1}^{k} Z_i \ge \mu k t\right\} \le \exp(k - kt/2) \ . \tag{2}$$

3 The Upper Bound

Lemma 1. *If G has diameter D then, with probability at least $1 - n^{1-K/2}$, the weight of the heaviest root-to-leaf path in the tree T is at most $K(\ln n + D)$, assuming that $K \ge 2$.*

Proof. Let v be a vertex of G such that there exists a path $P = v_0,\dots,v_k$ with $k \le D$ edges in G from $s = v_0$ to $v = v_k$. Let $e_i = v_{i-1}v_i$ be the ith edge on this path.

In Process FP, each edge e_i is assigned an exponential weight X_i. The path from s to v in T does not have length greater than $W(P) = \sum_{i=1}^{k} X_i$.

$$\begin{aligned}
\Pr\{W(P) \ge K(\ln n + D)\} &\le \Pr\{Z_1 + \cdots Z_k \ge K(\ln n + D)\} \\
&\le \exp(k - K(\ln n + D)/2) && \text{(using (2))} \\
&= n^{-K/2} \ .
\end{aligned}$$

For each $v \in V(G)$, let $W(v)$ denote the weight of the path, in T, from s to v, and define $W^* = \max\{W(v) : v \in V(G)\}$. as the weight of the longest root-to-leaf path in T. For each vertex v, G contains a path from s to v of length at most D. Therefore, by the discussion above and the union bond,so

$$\Pr\{W^* \ge K(\ln n + D)\} \le \sum_{v \in V(G)} \Pr\{t(v) \ge K(\ln n + D)\} \le n^{1-K/2} \ . \qquad \square$$

We are now ready to prove an upper bound on the height of T.

Theorem 1. *If G has diameter D and maximum degree Δ then,*

$$\Pr(h(T) \ge K\Delta e(2\ln n + D) = O(n^{-K/4}).$$

Proof. Let $P = v_0, v_1, \dots, v_L$ be some simple path in G starting at $s = v_0$ and let $e_i = v_{i-1}v_i$ be the ith edge on this path. In the model of Process FP, this path has a weight that is the sum of L independent exponential(1) random variables. Using (1), we obtain that for any $w > 0$ we have

$$\Pr\{W(P) \le w\} \le \left(\frac{ew}{L}\right)^L .$$

If P appears in T, then either $W(P) < 4\ln n + 2D$ or the heaviest root-to-leaf path in T has weight at least $4\ln n + 2D$. Therefore, by Lemma 1 and (1),

$$\Pr\{P \text{ appears in } T\} \le \frac{1}{n} + \left(\frac{e(4\ln n + 2D)}{L}\right)^L .$$

Let P_L be the set of all simple paths in G that start at s and have length L. The probability that T contains any path from P_L is at most

$$\frac{1}{n} + |P_L|\left(\frac{e(4\ln n + 2D}{L}\right)^L \le \frac{1}{n} + \left(\frac{\Delta e(4\ln n + 2D}{L}\right)^L ,$$

since $|P_L| \le \Delta^L$. Finally, the probability that T contains any path of length at least $L' = 2\Delta e(4\ln n + 2D)$ is at most

$$\frac{1}{n} + \sum_{L=L'}^{n-1}\left(\frac{\Delta e(4\ln n + 2D)}{L}\right)^L < \frac{1}{n} + \sum_{L=L'}^{n-1} 2^{-L} < \frac{2}{n} . \qquad \square$$

4 The Lower Bound

Note that Theorem 1 is obviously asymptotically tight for graphs with maximum degree $\Delta \in O(1)$ and diameter $D \in \Omega(\ln n)$. For instance, Theorem 1 shows that, for constant d, running Process A on a $n^{1/d} \times \cdots \times n^{1/d}$ grid or toroidal grid produces a spanning tree of diameter $\Theta(n^{1/d})$.

However, Theorem 1 is not tight for some graphs, like the d-dimensional hypercube, which has $n = 2^d$ vertices, maximum degree $\Delta = d/2 \in \Theta(\ln n)$ and diameter $D = d \in \Theta(\ln n)$. In this setting, a different argument shows that Process A produces a spanning tree of diameter $\Theta(\ln n)$ rather than the $O(\ln^2 n)$ bound given by Theorem 1.

It is natural to ask if one can eliminate or reduce the dependence on Δ in Theorem 1. Intuitively, higher degree should produce spanning trees of lower diameter. However, the example of a sequence of $D/2$ $(\Delta - 1)$-cliques joined in sequence (see Figure 1) shows that Process A may produce spanning trees of diameter $\Omega(D\ln\Delta)$. Therefore the dependence on Δ can not be eliminated entirely; there is at least a logarithmic dependence on Δ.

Next we show that, in fact, a linear dependence on Δ is necessary: For any Δ and any $D \in \Omega(\log\Delta)$, we can construct a graph $G = G_{D,\Delta}$ of diameter D and maximum degree Δ for which Process A produces a spanning tree of diameter $\Omega(\Delta D)$. Thus, the upper-bound of Theorem 1 is asymptotically tight for these graphs.

The graph G is obtained by gluing together two graphs H and I. The graph H has high diameter and high connectivity. The graph I has low connectivity and low diameter.

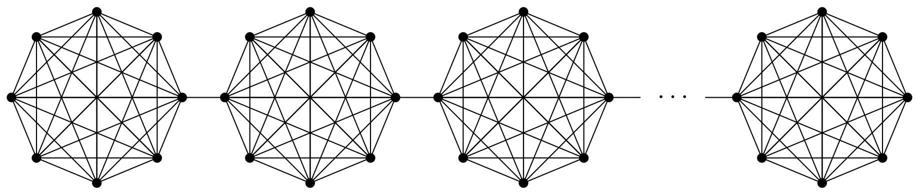

Figure 1: A sequence of $D/2$ $(\Delta-1)$-cliques gives a graph G with diameter D and maximum degree Δ for which Process A produces a spannning tree of height $\Omega(D \log \Delta)$.

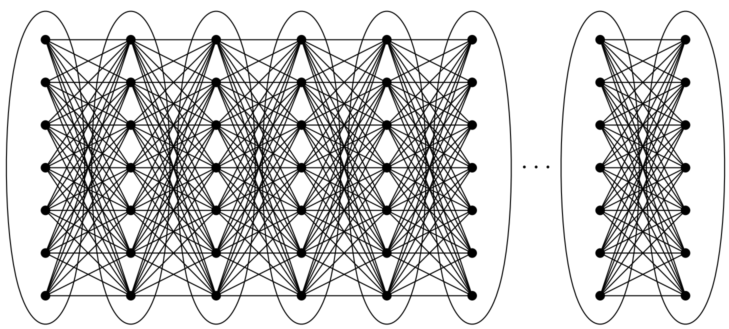

Figure 2: The graph H.

By joining them we obtain a graph of low diameter (because of I) but for which Process A is more likely to find paths in H (because of its high connectivity). We begin by defining and studying H and I independently.

4.1 The Ladder Graph H

Fix some integers $L, \delta \in \mathbb{N}$ to be described later and some constant $a > 1$, also described later. Refer to Figure 2. The vertices of H are partitioned in to L groups $V_1, \ldots, V_L$, each of size δ. The edge set of H is

$$E(H) = \bigcup_{i=1}^{L-1} \{vw : v \in V_i,\, w \in V_{i+1}\} \ .$$

First we show that H, under the models of Process E and Process FP has very low-weight paths between its vertices. Assign an independent exponential(1) weight to each edge of H. Let $d_H(v,w)$ denote the weight of the minimum weight path from v to w in the resulting weighted graph.

Lemma 2. *For any vertex $v \in V_i$ and any vertex $w \in V_j$, $j > i$,*

$$\Pr\{d_H(v,w) > t(j-i-1)/\delta + r)\} \le \begin{cases} \exp(-r) & \text{if } j-i=1 \\ \exp((1-t/2)(j-i-1)) + \exp(-r) & \text{otherwise.} \end{cases}$$

Proof. Consider the following greedy algorithm for finding a path from v to w: The path starts at v (which is in V_i). When the path has reached some vertex $x \in V_k$, for $k < j-1$, the algorithm extends the path by taking the minimum-weight edge joining x to some vertex in V_{k+1}. When the algorithm reaches some $x \in V_{j-1}$, it takes the edge xw.

Let $m = j-i$. Each of the first $m-1$ edges in the resulting path has a weight that is the minimum of δ exponential(1) random variables. Thus, the weight of these edges is the sum of $m-1$ exponential(δ) random variables $X_1,\ldots,X_{m-1}$. By (2),

$$\Pr\left\{\sum_{\ell=1}^{m-1} X_\ell > t(m-1)/\delta\right\} \le \exp((1-t/2)(m-1)) \ . \tag{3}$$

The last edge in this path has a weight X_m that is an exponential(1) random variable. From the definition of the exponential distribution,

$$\Pr\{X_m > r\} = \exp(-r) \ . \tag{4}$$

We complete the proof with the union bound:

$$\begin{aligned} \Pr\{d_H(v,w) > t(m-1)/\delta + r\} &= \Pr\left\{\sum_{\ell=1}^{m} X_m > t(m-1)/\delta + r\right\} \\ &\le \Pr\left\{\sum_{\ell=1}^{m-1} X_\ell > t(m-1)/\delta\right\} + \Pr\{X_m > r\} \\ &\le \exp((1-t/2)(m-1)) + \exp(-r) \ . \end{aligned}$$

□

Note that the proof of Lemma 2 actually studies the length of the greedy path from v to w; call this $d_H^{\mathrm{greedy}}(v,w)$. For a fixed k, $\Pr\{d_H^{\mathrm{greedy}}(v,w) > k\}$ is clearly maximized for $v \in V_1$ and $w \in V_L$. Therefore, by taking $r = aL/(e^2\delta)$ and $t = a/e^2$ (so that $tL/\delta + r = 2aL/(e^2\delta)$) we obtain the following special case of Lemma 2:

Corollary 1. *For any i and j and any $v \in V_i$, $w \in V_j$,*

$$\Pr\{d_H(v,w) > 2aL/(e^2\delta)\} \le \exp((1-a/(2e^2))L) + \exp(-aL/(e^2\delta)) \ .$$

4.2 The Subdivided Tree I

Next, we consider a graph I that is obtained by starting with a complete binary tree having L leaves and then subdividing each edge incident to a leaf $\lceil aL/\delta \rceil - 1$ times so that each leaf-incident edge becomes a path of length $\lceil aL/\delta \rceil$. Note that I has height $\lceil aL/\delta \rceil + \lceil \log_2 L \rceil$.

Assign independent exponential(1) edge weights to each edge of I and, for two leaves v and w, let $d_I(v,w)$ denote the weight of the unique path from v to w.

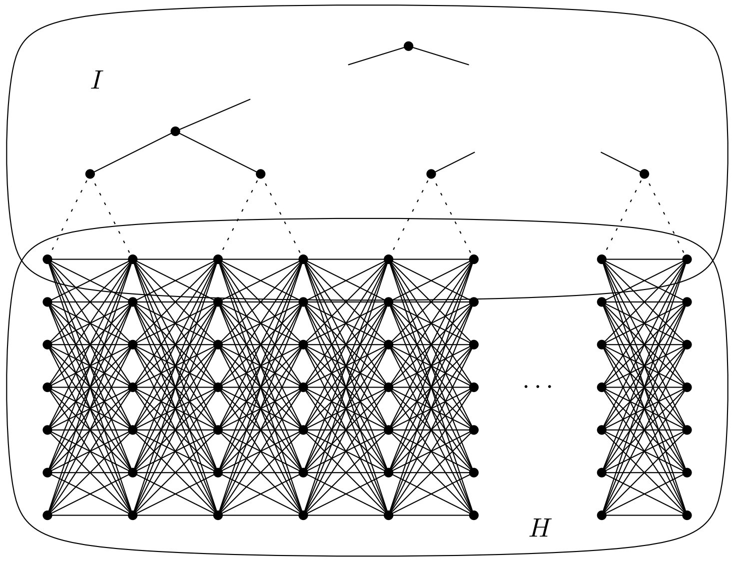

Figure 3: The lower bound graph G. Dotted segments denote subdivided edges (path of length $\lceil aL/\delta \rceil$).

Lemma 3. $\Pr\{d_I(v,w) \le 2aL/(e^2\delta)\} \le \exp(-2aL/\delta)$

Proof. The path from v to w in I contains at least $2\lceil aL/\delta \rceil$ edges. Therefore, the weight of this path is lower-bounded by the sum of $2\lceil aL/\delta \rceil$ independent exponential(1) random variables. The lemma then follows by applying (1) to this sum. □

4.3 Putting it Together

The lower-bound graph G is now constructed by taking a tree I with L leaves and a graph H with L groups $V_1, \ldots, V_L$ each of size $\delta = \lfloor (\Delta - 1)/2 \rfloor$. Next, we consider the leaves of I in the order they are encountered in a depth first-traversal of I and, for each $i \in \{1, \ldots, L\}$ we identify the ith leaf of I with some vertex in V_i. See Figure 3.

Note that the graph G has maximum degree $\Delta \le 2\delta + 1$. Furthermore, every vertex of G is either in I, or adjacent to a vertex in I. Therefore, G has diameter

$$D = 2 + 2(\ln L + aL/\delta) = O(L/\Delta) \ ,$$

for $L \in \Omega(\Delta \ln \Delta)$.

Note that the graph G has three parameters a, L, and Δ, so we will call this graph $G(a, L, \Delta)$.

Theorem 2. *For every $\Delta \ge 3$ and every $L \in \Omega(\Delta \ln \Delta)$, there exists a constant a such that If we run Process A on $G(a, L, \delta)$ starting at some vertex $s \in V_1$, then with probability at least $1 - o_L(1)$, the resulting spanning tree contains a path of length at least $L - 1$.*

Proof. In the Process FP view, we assign each edge of G an exponential(1) edge weight and compute a shortest path tree T rooted at s in the resulting weighted graph. Consider the path P in T from a vertex s in V_1 to an arbitrary vertex t in V_L. If P uses no edges of I, then P uses at least $L-1$ edges. If P does use some edge of I, then this implies that there are two leaves v and w of I such that $d_G(s,t) \geq d_H(v,w) \geq d_I(v,w)$.

Using Corollary 1 and Lemma 3, we have

$$\begin{aligned}\Pr\{d_H(v,w) \geq d_I(v,w)\} &\leq \binom{L}{2}\Big(\Pr\{d_H(v,w) > 2aL/e^2\delta\} + \Pr\{d_I(v,w) < 2aL/e^2\delta\}\Big) \\ &\leq \binom{L}{2}\Big(\exp((1-a/2e^2)L) + \exp(-aL/(e^2\delta)) + \exp(-2aL/\delta)\Big)\end{aligned}$$

This probability tends to zero as L grows, provided that $a \geq \max\{4e^2, 3e^2\delta \ln L/L\}$. Such a constant a exists whenever $\delta \ln L/L = O(1)$ i.e. for any $L \in \Omega(\Delta \log \Delta)$. □