ON THE DISCOVERY OF THE SEED IN UNIFORM ATTACHMENT TREES

LUC DEVROYE AND TOMMY REDDAD

Abstract. We investigate the size of vertex confidence sets for including part of (or the entirety of) the seed in seeded uniform attachment trees, given knowledge of some of the seed's properties, and with a prescribed probability of failure. We also study the problem of identifying the leaves of a seed in a seeded uniform attachment tree, given knowledge of the positions of all internal nodes of the seed.

1. Introduction

In the web graph, nodes represent websites, and edges represent links. Its theoretical study requires proper graph models that explain and approximate what is observed in the field—see [5] for an early book on this topic. One particular set of models grows the web graph dynamically, one new website at a time. This leads to the study of random graph dynamics [15, 28]. The main—but still simple—models here are the uniform random recursive tree and the preferential attachment model. Simple generalizations of these tree models in which k instead of one parent are selected at each step lead to graphs. We are concerned in this paper with the estimation of the origin of the tree when one is shown the entire (unrooted) tree. In particular, we consider only uniform attachment trees that are grown started from a fixed tree S, called the seed, a problem first studied by Bubeck *et al*. [8, 9]. Estimating the seed can aid, for example, in the identification of the source of a rumour, an idea, or an epidemic. Finding the source of an epidemic can help to identify the original causes for the spread of the infection, and aid in the development of preventative measures. The area of discovering the origins of the web graph or a social network graph is also called *network archeology.*

Consider the random tree which begins as some fixed tree S and grows incrementally by adding a new vertex and connecting it to a uniformly random node among all nodes. We think of the starting tree as the *seed* of the process, and we continue the attachment process for a long time. We ask the following question: How difficult is it to identify the position of the seed in such a process only given the structure of the tree? In general, what aspects of the seed can be identified efficiently?

The growth process, started from a single node, yields the uniform random recursive tree (URRT) or uniform attachment tree (see *e.g.,* [12, 14, 23, 24]). We take the following point of view, following [7]: Given the uniform attachment tree T, and a fixed $\varepsilon > 0$, our algorithm returns a set $H = H(T, \varepsilon)$ of nodes such that,

Date: June 25, 2019.

2010 *Mathematics Subject Classification.* Primary: 05C80.

Key words and phrases. Random trees, seed tree, root finding, uniform attachment.

in the case that $|S| = 1$,

$$\mathbf{P}\{V(S) \subseteq H\} \geq 1 - \varepsilon,$$

where $V(S)$ denotes the set of vertices of S. That this is even possible regardless of the size of T is interesting. In [7], it is shown that there exist universal constants $c_1, c_2 > 0$ such that for any algorithm,

$$|H| \geq c_1 \exp\Big\{c_2 \sqrt{\log(1/\varepsilon)}\Big\}.$$

Furthermore, an algorithm is given in [7] that has for some other universal constants $c_1, c_2 > 0$,

$$|H| \leq c_1 \exp\left\{c_2 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\right\}.$$

Consider now seeds S with k vertices and ℓ leaves, where k and ℓ are known. We study algorithms that return sets H or H^*, both depending upon k, ℓ, and ε, having the properties that

$$\mathbf{P}\{V(S) \subseteq H\} \geq 1 - \varepsilon,$$

and

$$\mathbf{P}\{|V(S) \cap H^*| \geq 1\} \geq 1 - \varepsilon,$$

respectively. To be a bit more formal, if we write $L(T)$ for the set of leaves of the tree T, $A^{(m)}$ to be the set of all m-sized subsets of the set A, $\mathcal{T}$ for the space of all unlabelled trees, and $V(\mathcal{T})$ for the set of all vertices in these trees, we consider algorithms with input T (our tree), k, ℓ, and ε, and with K-sized set-valued output, and introduce the optimal sizes

$$K(k, \ell, \varepsilon) = \min\left\{ m\colon \begin{array}{c} \exists H_{m,k,\ell,\varepsilon}\colon \mathcal{T} \to V(\mathcal{T})^{(m)}, \text{ such that} \\ \displaystyle\min_{\substack{S\colon |S|=k \\ |L(S)|=\ell}} \mathbf{P}\{V(S) \subseteq H_{m,k,\ell,\varepsilon}(T)\} \geq 1 - \varepsilon \end{array} \right\},$$

$$K^*(k, \ell, \varepsilon) = \min\left\{ m\colon \begin{array}{c} \exists H^*_{m,k,\ell,\varepsilon}\colon \mathcal{T} \to V(\mathcal{T})^{(m)}, \text{ such that} \\ \displaystyle\min_{\substack{S\colon |S|=k \\ |L(S)|=\ell}} \mathbf{P}\{|V(S) \cap H^*_{m,k,\ell,\varepsilon}(T)| \geq 1\} \geq 1 - \varepsilon \end{array} \right\}.$$

In the case that $k = 1$, the seed is a single node, and we simply write $K(\varepsilon) = K(1, 1, \varepsilon)$.

We can also introduce $K(S, \varepsilon)$ and $K^*(S, \varepsilon)$, the analogous quantities in which the full structure of S is assumed, where we now have

$$K(S, \varepsilon) = \min\Big\{m\colon \exists H_{m,S,\varepsilon}\colon \mathcal{T} \to V(\mathcal{T})^{(m)}, \mathbf{P}\{V(S) \subseteq H_{m,S,\varepsilon}(T)\} \geq 1 - \varepsilon\Big\},$$

$$K^*(S, \varepsilon) = \min\Big\{m\colon \exists H^*_{m,S,\varepsilon}\colon \mathcal{T} \to V(\mathcal{T})^{(m)}, \mathbf{P}\{|V(S) \cap H^*_{m,S,\varepsilon}(T)| \geq 1\} \geq 1 - \varepsilon\Big\}.$$

Our main results are as follows:

Theorem 1.1. *There are universal constants $c, \varepsilon_0 > 0$ such that if $\varepsilon \leq \varepsilon_0$, then*

$$K^*(k, \ell, \varepsilon) \leq c(1/\varepsilon)^{2/k} \log(1/\varepsilon).$$

The following result shows that for a fixed k, the dependence of $K^*(k, \ell, \varepsilon)$ is at most subpolynomial in $1/\varepsilon$.

Theorem 1.2. *There are universal constants $c_1, c_2, c_3, c_4 > 0$ such that for $k > c_1$ and $\varepsilon \leq \exp\{-c_2(\log k)^{11}\}$,*

$$K^*(k,\ell,\varepsilon) \leq c_3 \exp\left\{c_4 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon) + \log\log k}\right\}.$$

We note that the bound in Theorem 1.1 is better than that of Theorem 1.2 if k is much larger than $\log\log(1/\varepsilon)$.

Theorem 1.3.

$$K^*(k,\ell,\varepsilon) \geq K((e\varepsilon)^{1/k}).$$

As an immediate corollary, in view of the lower bound of Bubeck, Devroye, and Lugosi [7, Theorem 4] on $K(\varepsilon)$, we have the following explicit lower bound.

Corollary 1.4. *There are universal constant $c_1, c_2, c_3 > 0$ such that for all $\varepsilon \leq e^{-c_1 k}$,*

$$K^*(k,\ell,\varepsilon) \geq c_2 \exp\left\{c_3 \sqrt{\frac{\log(1/\varepsilon)}{k}}\right\}.$$

It should be noted that the above four bounds do not depend on the value of ℓ.

Theorem 1.5. *There are universal constants $c, \varepsilon_0 > 0$ such that if $\varepsilon \leq \varepsilon_0$, then*

$$K(k,\ell,\varepsilon) \leq (ck\ell/\varepsilon) \min\Big\{\log(k\ell/\varepsilon), K^*(k,\ell,\varepsilon/2)\Big\}.$$

In conjunction with Theorem 1.1,

$$K(k,\ell,\varepsilon) \leq (ck\ell/\varepsilon) \min\Big\{\log(k\ell/\varepsilon), (1/\varepsilon)^{2/k}\log(1/\varepsilon)\Big\}.$$

We also study the optimal size $K'(k,\ell,\varepsilon)$, which is the size of the smallest set of vertices to include all leaves of a seed S with $|S| = k$ and $|L(S)| = \ell$, given we already know the position of its internal nodes. In this case, the dependence of $K'(k,\ell,\varepsilon)$ is shown to be only logarithmic in $1/\varepsilon$, in contrast to the preceding results.

Theorem 1.6. *There are universal constants $c, \varepsilon_0 > 0$ for which, if $\varepsilon \leq \varepsilon_0$, then*

$$K'(k,\ell,\varepsilon) \leq \ell + c(k-\ell)\log\left((\ell/\varepsilon)\log\left(\frac{k-\ell}{\varepsilon}\right)\right).$$

Proposition 4.3 shows that Theorem 1.6 is tight for a large class of seeds.

We also prove that assuming knowledge of the full structure of the seed can make things much easier.

Theorem 1.7. *There are universal constants $c, \varepsilon_0 > 0$ such that if $\varepsilon \leq \varepsilon_0$, then*

$$K(S_k,\varepsilon) \leq c(k + \log(1/\varepsilon))(1/\varepsilon)^{1/k}\log(1/\varepsilon),$$

where S_k is a star on k vertices.

Some of the above quantities can be related using the following simple inequalities which we state without proof:

$$K^*(k,\ell,\varepsilon) \leq K(k,\ell,\varepsilon); \tag{1}$$

$$(k-\ell) + K'(k,\ell,\varepsilon) \leq K(k,\ell,\varepsilon); \tag{2}$$

$$K(S,\varepsilon) \leq K(|S|,|L(S)|,\varepsilon). \tag{3}$$

1.1. **Related work.** The oldest work on this topic seems to be by Haigh [16], who in 1969 studied properties of the maximum likelihood estimate of the root in a uniform attachment tree, including the precise identification of the limiting probability $1 - \log 2$ of success when only one candidate node can be selected. Shah and Zaman [25] studied properties of the maximum likelihood estimate of the root in a diffusion process over regular trees [25]. A *diffusion process* over an infinite graph G is a sequence $(G_1, G_2, \dots)$ described by designating a root vertex u_1, where $G_1 = \{u_1\}$, and G_{i+1} is obtained from G_i by adding to G_i a uniformly random edge in its boundary. Shah and Zaman defined a measure of node centrality called *rumor centrality* which they showed coincided with the maximum likelihood estimate for the root of a diffusion process in a regular tree. In a follow-up work [26], Shah and Zaman study the effectiveness of rumor centrality as an estimator of the root for a larger family of random trees. Bubeck, Devroye, and Lugosi [7] independently studied root-finding in uniform and preferential attachment trees. They showed that rumor centrality could serve as an effective estimator for the root in a uniform attachment tree, and gave explicit bounds on the size of vertex-confidence sets for the root. It is also shown in [7] that root-finding is possible in preferential attachment trees, for instance by picking nodes of high degree. Jog and Loh [17] showed that root-finding algorithms also exist for sublinear preferential attachment trees. Khim and Loh [20] gave bounds on the size of vertex-confidence sets for the root in a diffusion process over regular trees. They also showed that root-finding is possible over a certain asymmetric infinite tree. In general, diffusion processes are part of the study of first-passage percolation, which is an old and widely-studied subject. For a recent survey of classical and newer works in this field, see [1].

Every root-finding algorithm discussed in the above works depends upon measures of node centrality. In the uniform and preferential attachment models, as well as in a diffusion process over d-ary trees, Jog and Loh also showed that with high probability, a single node persists as the most central node throughout the process, after a finite number of steps [18].

Seeded attachment trees have received some more attention lately, where Bubeck, Mossel, and Rácz [9] first showed that the total variation distance between between the distributions of arbitrarily large seeded preferential attachment trees is lower bounded by a positive constant, as long as the two seeds have distinct degree profiles. This result was later extended to hold for all non-isomorphic pairs of seeds by Curien, Duquesne, Kortchemski, and Manolescu [10], and further modified to work for uniform attachment trees by Bubeck, Eldan, Mossel, and Rácz [8]. The influence of the seed in either sublinear or superlinear preferential attachment trees remains unknown.

Recently, Lugosi and Pereira [21] specifically studied the problem of partially recovering the seed in seeded uniform attachment trees in which the seed is either a path, a star, or a uniform attachment tree itself. In particular, they show that there are universal constants $c_1, c_2 > 0$ such that if $k \geq c_1 \log(1/\varepsilon)$, then $K(S_k, \varepsilon) \leq c_2 k$ [21, Theorem 3]. In comparison to our Theorem 1.7, we note that our result works for all values of k, while their result gives a better joint dependence on k and ε in the applicable range.

1.2. **Content of the paper.** To start in Section 2, we focus on algorithms designed to report sets of vertices which intersect the seed with probability at least $1 - \varepsilon$. In Section 2.1, we use basic results about Pólya urns to give a simple algorithm which reports a set of nodes which are known to intersect the seed with probability at least $1-\varepsilon$. This gives an upper bound on $K^*(k,\ell,\varepsilon)$. In Section 2.3, we study a different algorithm which gives an improvement on the dependence of $K^*(k,\ell,\varepsilon)$ on $1/\varepsilon$, for a certain range of k and ε. We also leverage a lower bound from [7] to give a lower bound on $K^*(k,\ell,\varepsilon)$ in Section 2.4, which ultimately relies upon the maximum likelihood estimate for root-finding vertex confidence sets.

In Section 3, our focus shifts to the analysis of algorithms which return sets which include the entire seed with probability at least $1-\varepsilon$. Using a slightly different analysis of the same algorithm as in Section 2.1, we give the first upper bound on $K(k,\ell,\varepsilon)$ from Theorem 1.5 in Section 3.1. We give yet another simple algorithm to solve this problem, which relies on the upper bound on $K^*(k,\ell,\varepsilon)$ of Section 2.1, thereby proving the second part of Theorem 1.5.

In Section 4, we investigate algorithms for locating all nodes of the seed when we assume knowledge of the positions of all internal nodes. We further show in Section 5 that if we know that the seed is a star on k nodes, then only a constant factor of k nodes suffice to identify all nodes of the seed with constant probability.

1.3. Notation and background.

For a set A and $i \in \mathbb{N}$, let $A^{(i)}$ denote the set of all i-sized subsets of A, *i.e.,* $A^{(i)} = \{B \colon |B| = i, B \subseteq A\}$. Let also $\log\colon (0,\infty) \to \mathbb{R}$ denote the natural logarithm $\log_e$.

Throughout this document, we write S for a tree with $V(S) = \{u_1,\dots,u_k\}$ and with ℓ leaves. The leaf set of S is denoted by $L(S)$. In general, we write $|S|$ instead of $|V(S)|$ for the number of vertices in a tree, so $|S| = k$. The tree S is called the *seed*, and we assume throughout that $k \geq 2$.

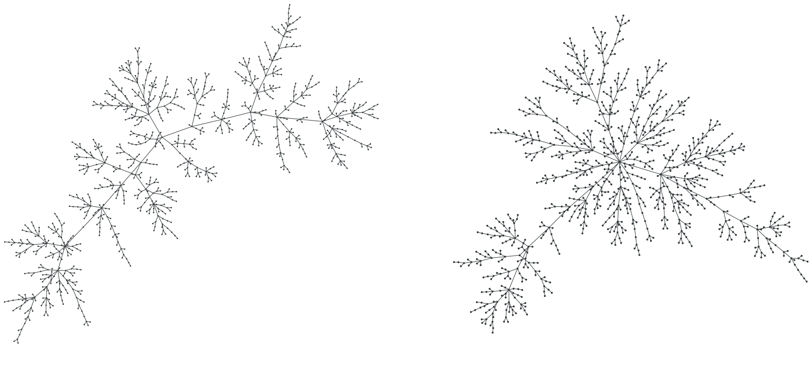
For a tree T, we say that T' is a *subtree* of T if the vertices of T' induce a connected subgraph of T. For T with subtree T' and $u \in V(T)$, let $(T,T')_{u\downarrow}$ and $(T,V(T'))_{u\downarrow}$ denote the subtree of T rooted at u "facing away" from T', *i.e.,* $(T,T')_{u\downarrow}$ is the subtree of T induced on all nodes whose (unique) path to T' includes the vertex u.

For a tree T and set $X \subseteq V(T)$, we write $N_T(X)$ for the set of neighbours of all vertices in X, *i.e.,*

$$N_T(X) = \{u \in V(T) \setminus X \colon u \text{ is adjacent to } v \text{ for some } v \in X\}.$$

With a slight abuse of notation, we will write $\deg_T(X) = |N_T(X)|$, and for $u \in V(T)$, $\deg_T(u) = \deg_T(\{u\})$. We will omit the subscript in $\deg_T$ and N_T whenever the tree T is understood.

We inductively define a distribution on labelled trees: Let $\alpha \geq 0$, let $\mathrm{UA}_\alpha(k,S) = S$, and suppose that we are given $T_{n-1} \sim \mathrm{UA}_\alpha(n-1,S)$, where $V(T_{n-1}) = \{u_1,\dots,u_{n-1}\}$. Let $T_n \sim \mathrm{UA}_\alpha(n,S)$ be obtained from T_{n-1} by adding a leaf labelled u_n to T_{n-1} and connecting it to a vertex $u \in \{u_1,\dots,u_{n-1}\}$ with probability proportional to $\deg_{T_{n-1}}(u)^\alpha$. The distribution $\mathrm{UA}_0(n,P_2)$, in which new leaves are attached to uniformly random vertices sequentially, is called the *uniform attachment tree* or *random recursive tree*, and is sometimes denoted by $\mathrm{UA}(n)$ or $\mathrm{URRT}(n)$ in the literature. The distribution $\mathrm{UA}_1(n,P_2)$ is called the *(linear) preferential attachment tree* or *Barabási-Albert model* [2], sometimes denoted by $\mathrm{PA}(n)$ in the literature. For $\alpha \in (0,1)$, $\mathrm{UA}_\alpha(n,P_2)$ is called the *sublinear preferential attachment tree*, and when $\alpha > 1$, it is called the *superlinear preferential attachment tree.* For general seeds S, the distributions above are said to be *with seed S*. In this paper, we are mostly concerned with $\mathrm{UA}_0(n,S)$, so we generally omit the subscript



(A) S is a path on 10 nodes.

(B) S is a star on 10 nodes.

FIGURE 1. One sample from $\mathrm{UA}(1000, S)$ for two different seeds.

"0." Moreover, if $T \sim \mathrm{UA}(n, S)$ and the distribution of T is understood, we avoid repeating its distribution. See Figure 1 for a typical sample from $\mathrm{UA}(n, S)$ for different choices of S.

The vertices of the aforementioned trees are labelled; for a (rooted or unrooted) labelled tree T, let T° denote the isomorphism class of T, *i.e.,* the operation $\circ$ "forgets" the labelling of T. With some abuse of notation, we refer to nodes of T° using their original labels in T. Of course, if $T \sim \mathrm{UA}(n, S)$ and nothing at all is assumed about S, it is always possible that $S = T$. Therefore, except if otherwise specified, we assume knowledge of the size of the seed in trying to locate it in T°—our main goal is to detect S in the unlabelled tree $\mathrm{UA}(n, S)^\circ$, given that $|S| = k$.

In general, sets of vertices which are made to include parts of the seed are referred to as *(root-finding) vertex confidence sets.* We reserve the letter H for functions which map unlabelled trees to vertex confidence sets, and such functions are called *root-finding algorithms.* We also reserve the letter K for the size of vertex confidence sets.

We assume that the reader is familiar with the basics of probability theory, including basic properties of standard random variables. We write $\mathrm{Beta}(\alpha, \beta)$ for a Beta distribution with parameters α, β, *i.e.,* the distribution supported on $[0, 1]$ with density

$$f_{\alpha,\beta}(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1},$$

where for $(\alpha_1, \ldots, \alpha_k) \in \mathbb{R}_+^k$, we have $\mathrm{B}(\alpha_1, \ldots, \alpha_k) = \frac{\prod_{i=1}^k \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i)}$. We write $\mathrm{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ for the Dirichlet distribution, which is supported on the k-simplex

$$\left\{ (x_1, \ldots, x_k) \colon \sum_{i=1}^k x_i = 1 \text{ and } x_i \in [0, 1] \right\}$$

and which has density

$$f_{\alpha_1,\dots,\alpha_k}(x_1,\dots,x_k) = \frac{1}{\mathrm{B}(\alpha_1,\dots,\alpha_k)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

We often make use of the following fact about the distribution of sums of Dirichlet marginals: If $(X_1,\dots,X_k) \sim \mathrm{Dirichlet}(\alpha_1,\dots,\alpha_k)$, and $I \subseteq \{1,\dots,k\}$ is some index set, then

$$\sum_{i \in I} X_i \sim \mathrm{Beta}\left(\sum_{i \in I} \alpha_i, \sum_{i \in \{1,\dots,k\} \setminus I} \alpha_i \right).$$

1.4. **Acknowledgements.** The authors would like to thank the anonymous referees for their helpful comments and suggestions. Luc Devroye is supported by NSERC Grant A3456. Tommy Reddad is supported by an NSERC PGS D scholarship 396164433.

2. Finding a single vertex in the seed

2.1. **An upper bound.** For a tree T and $u \in V(T)$, let

$$\psi_T(u) = \max_{v \in V(T) - u} |(T,u)_{v\downarrow}|,$$

so $\psi_T(u)$ is the size of the largest subtree of T hanging off of the vertex u. We omit the subscript T when the tree is understood.

Consider the strategy for picking central nodes from an unlabelled copy of $T \sim \mathrm{UA}(n, S)$ which works by simply picking the K nodes minimizing the values of ψ. We will write $H^*_{\psi;K}(T^\circ)$ for such a set of nodes. This strategy was first successfully introduced in [7], where ψ is seen as a measure of node centrality, and in which it is shown that the root of $\mathrm{UA}(n)$ (which is the node labelled u_1) is likely to have low ψ value. In fact, it is shown that this is also true for the root in preferential attachment trees. This centrality measure is further studied for root-finding in different settings [17, 18, 20].

Specifically, the result of [7, Theorem 3] has that for $K \geq (2.5/\varepsilon)\log(1/\varepsilon)$,

$$\mathbf{P}_{T \sim \mathrm{UA}(n)} \left\{ u_1 \in H^*_{\psi;K}(T^\circ) \right\} \geq 1 - \frac{4\varepsilon}{1-\varepsilon},$$

and so, for $\varepsilon \leq 1/2$, $K(\varepsilon) \leq (80/\varepsilon)\log(1/\varepsilon)$. The following result indicates that $H^*_{\psi;K}$ also serves as a root-finding algorithm for seeded uniform attachment seeds, and offers an upper bound on the size of such vertex confidence sets.

Our proof relies on a few supporting lemmas and a classical result in the study of Pólya urns, whose proofs and statements are given in Appendix A.

Proposition 2.1. *There are universal constants $c, \varepsilon_0 > 0$ for which, if $\varepsilon \leq \varepsilon_0$ and*

$$K \geq c(1/\varepsilon)^{2/k} \log(1/\varepsilon),$$

then

$$\mathbf{P}\{|V(S) \cap H^*_{\psi;K}(T^\circ)| \geq 1\} \geq 1 - \varepsilon.$$

Theorem 1.1 follows immediately.

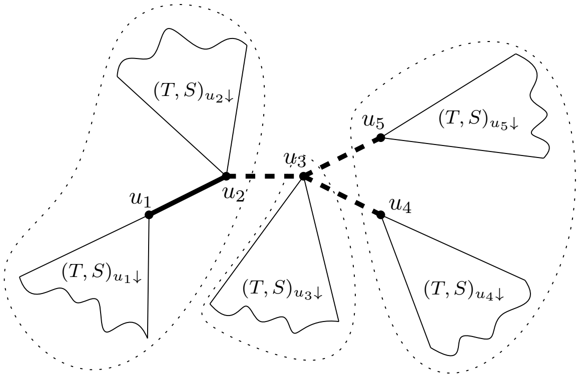


FIGURE 2. The seed S has vertex set $\{u_1,\ldots,u_5\}$. The node $r = u_3$ is its unique centroid. The dashed edges form the set $\partial_S(r)$. The dotted lines outline the components of $S - \partial_S(r)$.

Proof of Proposition 2.1. Let $\psi_* = \min\{\psi(u_1),\ldots,\psi(u_k)\}$. If for all $i > K$, $\psi(u_i) > \psi_*$, then $V(S)$ intersects $H^*_{\psi;K}(T^\circ)$, so

$$\begin{aligned}\mathbf{P}\{V(S)\cap H^*_{\psi;K}(T^\circ)=\emptyset\} &\le \mathbf{P}\{\exists i>K\colon \psi(u_i)\le\psi_*\}\\ &\le \mathbf{P}\{\psi_*\ge nt\}+\mathbf{P}\{\exists i>K\colon \psi(u_i)\le nt\},\end{aligned}\tag{4}$$

where $t > 0$ is to be specified later. It is a classical result that the seed S has a centroid, *i.e.,* a node r whose removal splits the seed into components each of size at most $k/2$ [19]. Note that

$$\psi_* \le \psi(r) \le \max_{C\in\mathcal{C}(S-\partial_S(r))} \sum_{u\in C} |(T,S)_{u\downarrow}|,$$

where $\mathcal{C}(G)$ denotes the set of components of a graph G, and $\partial_S(r)$ denotes all of the edges of S incident to r. Now, for any $C \in \mathcal{C}(S-\partial_S(r))$, we have by Lemma A.2 that

$$\frac{1}{n}\sum_{u\in C}|(T,S)_{u\downarrow}| \xrightarrow{d} \mathrm{Beta}(|C|, k-|C|),$$

as $n\to\infty$, so stochastically,

$$\frac{\psi_*}{n} \le \max_{C\in\mathcal{C}(S-\partial_S(r))} \mathrm{Beta}(|C|, k-|C|)$$

and since $|C|\le k/2$ uniformly because r is a centroid, then each such beta random variable is stochastically smaller than a $\mathrm{Beta}(k/2,k/2)$ by Lemma A.3. See Figure 2. Let $f_{k/2,k/2}(x)$ be the density of a $\mathrm{Beta}(k/2,k/2)$ random variable. Using the bound

$$\sqrt{\frac{2\pi}{x}}\left(\frac{x}{e}\right)^x \le \Gamma(x) \le \sqrt{\frac{2\pi}{x}}\left(\frac{x}{e}\right)^x e^{\frac{1}{12x}},$$

which holds for all $x > 0$ [13, Equation 5.6.1], we see that

$$\mathrm{B}(k/2,k/2) = \frac{\Gamma(k/2)^2}{\Gamma(k)} \ge 2^{-k}\cdot\frac{2}{e^{1/12}}\sqrt{\frac{2\pi}{k}} > 3^{-k}$$

for all $k \geq 1$. Then,

$$\begin{aligned}
\mathbf{P}\{\psi_* \geq nt\} &\leq k\,\mathbf{P}\{\mathrm{Beta}(k/2, k/2) \geq t\} \\
&= k \int_t^1 f_{k/2,k/2}(x)\,\mathrm{d}x \\
&\leq k3^k \int_0^{1-t} (x(1-x))^{k/2-1}\,\mathrm{d}x \\
&\leq k3^k \int_0^{1-t} x^{k/2-1}\,\mathrm{d}x \\
&= 2(1-t)^{k/2}3^k,
\end{aligned}$$

so we can pick $t = 1 - (1/9)(\varepsilon/4)^{2/k}$.

To summarize, we have shown that we can bound the first term in (4) by

$$\mathbf{P}\{\psi_*/n \geq 1 - (1/9)(\varepsilon/4)^{2/k}\} \leq \varepsilon/2.$$

For the second term, the argument is identical to that of [7, Theorem 3]. Indeed, if T_i denotes the subgraph of T containing the vertex u_i after the removal of all edges between $\{u_1, \ldots, u_K\}$, then for any $i > K$,

$$\psi(u_i) \geq \min_{1 \leq j \leq K} \sum_{m=1, m\neq j}^{K} |T_m|,$$

and by Lemma A.2,

$$\frac{1}{n} \sum_{m=1, m\neq j}^{K} |T_m| \xrightarrow{d} \mathrm{Beta}(K-1, 1)$$

as $n \to \infty$, so that

$$\begin{aligned}
\mathbf{P}\{\exists i > K \colon \psi(u_i) \leq nt\} &\leq K\,\mathbf{P}\{\mathrm{Beta}(K-1, 1) \leq t\} \\
&= Kt^{K-1} \\
&\leq Ke^{-(1/9)\varepsilon^{2/k}(K-1)},
\end{aligned}$$

and a little bit of arithmetic shows that we should pick

$$K \geq c(1/\varepsilon)^{2/k} \log(1/\varepsilon)$$

for universal constants $c, \varepsilon_0 > 0$, as long as $\varepsilon \leq \varepsilon_0$. □

Consequently, we see that

$$K^*(k, \ell, \varepsilon) \leq c(1/\varepsilon)^{2/k} \log(1/\varepsilon).$$

We note that $H^*_{\psi;K}(T^\circ)$ can be computed in $O(n)$ time.

2.2. The maximum likelihood estimate.

Given an unlabelled tree T, and a candidate seed S, define the likelihood function $\mathcal{L}_T(S)$ to be the probability of observing T under $\mathrm{UA}(n, S)^\circ$, *i.e.*,

$$\mathcal{L}_T(S) = \mathbf{P}_{T' \sim \mathrm{UA}(n,S)}\{T'^\circ = T\},$$

and if $\mathcal{S}_{k,\ell}(T)$ denotes the set of all possible seeds in T with k vertices and ℓ leaves, the *maximum likelihood estimate* for S is given by

$$S^* = \operatorname*{argmax}_{S \in \mathcal{S}_{k,\ell}(T)} \mathcal{L}_T(S).$$

Note that the subtrees $(T, S)_{u\downarrow}$ for $u \in V(S)$ are, conditionally on their sizes, independent random recursive trees:

Lemma 2.2. *Let S be some seed, and $T \sim \mathrm{UA}(n, S)$, and $n_u \in \mathbb{N}$ for $u \in S$ be such that $\sum_{u\in S} n_u = n$. Then, conditionally on $|(T, S)_{u\downarrow}| = n_u$ for all $u \in S$, the trees $(T, S)_{u\downarrow}$ are independently distributed as $\mathrm{UA}(n_u)$.*

Proof sketch. Any incoming node in the attachment process $\mathrm{UA}(n, S)$, conditionally upon connecting to $(T, S)_{u\downarrow}$ for some $u \in S$, will connect to a uniformly random node of $(T, S)_{u\downarrow}$. Conditioning on the event that $|(T, S)_{u\downarrow}| = n_u$, this precisely describes the uniform attachment tree $\mathrm{UA}(n_u)$. □

As a consequence,

$$\mathcal{L}_T(S) = \prod_{u \in V(S)} \mathcal{L}_{(T,S)_{u\downarrow}}(u), \tag{5}$$

where $\mathcal{L}_{(T,S)_{u\downarrow}}(u)$ is the likelihood of the tree $(T, S)_{u\downarrow}$ to be rooted at u, which is computed in [7, Section 3]. Specifically, as in [7], let T be a rooted tree and v a vertex of T, and $T_1, \ldots, T_k$ be the subtrees rooted at the children of v listed in an arbitrary order, and $S_1, \ldots S_L$ be the isomorphism classes realized by these subtrees, define

$$\mathrm{Aut}(v, T) = \prod_{i=1}^{L} \bigl|\{j \in \{1, \ldots, k\} \colon T_j^\circ = S_i\}\bigr|!.$$

Define also, for an unrooted unlabelled tree T,

$$\overline{\mathrm{Aut}}(u, T) = |\{v \in V(T) \colon (T, v)^\circ = (T, u)^\circ\}|,$$

i.e., $\overline{\mathrm{Aut}}(u, T)$ is the number of vertices $v \in V(T)$ such that the rooted trees (T, v) and (T, u) are isomorphic. Then,

$$\mathcal{L}_T(u) = \frac{|T|}{\overline{\mathrm{Aut}}(u, T)} \prod_{v \in V(T)} \frac{1}{|(T, u)_{v\downarrow}| \,\mathrm{Aut}(v, (T, u))},$$

so

$$\mathcal{L}_T(S) = \prod_{u \in S} \frac{|(T, S)_{u\downarrow}|}{\overline{\mathrm{Aut}}(u, (T, S)_{u\downarrow})} \prod_{v \in (T,S)_{u\downarrow}} \frac{1}{|(T, S)_{v\downarrow}| \,\mathrm{Aut}(v, (T, S)_{u\downarrow})}. \tag{6}$$

In particular, this implies that the maximum likelihood estimate S^* can be computed in polynomial time in n for any fixed k, ℓ, just as in the case when $k = 1$ [7].

2.3. A vertex-confidence set with size subpolynomial in $(1/\varepsilon)$.

For a tree T and $u \in V(T)$, let

$$\varphi_T(u) = \prod_{v \in V(T) - \{u\}} |(T, u)_{v\downarrow}|,$$

omitting the T subscript when T is understood. We note, as noted by Bubeck, Devroye, and Lugosi in [7], that $1/\varphi$ resembles the expression of the likelihood in (6), so that nodes with small values of φ should be likely candidates for the root. For $K \geq 1$, let $H^*_{\varphi;K}(T^\circ)$ denote the set of K vertices in T° minimizing their values of φ. In [7], it is shown that $H^*_{\varphi;K}$ serves as an effective root-finding algorithm for $T \sim \mathrm{UA}(n)$. We show that this is also the case when $T \sim \mathrm{UA}(n, S)$.

Proposition 2.3. *There are universal constants $c_1, c_2, c_3, c_4 > 0$ such that if $k > c_1$, $\varepsilon \leq \exp\{-c_2(\log k)^{11}\}$, and*

$$K \geq c_3 \exp\left\{c_4 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon) + \log\log k}\right\},$$

then

$$\mathbf{P}\{|V(S) \cap H^*_{\varphi;K}(T^\circ)| \geq 1\} \geq 1 - \varepsilon.$$

Proof. Let u be a vertex of S, and label each node $v \in (T, S)_{u\downarrow}$ using an extended version of the labelling scheme from [7], where every node of T gets a label in

$$\mathbb{N}^* = \mathbb{N} \cup \mathbb{N}^2 \cup \mathbb{N}^3 \cup \dots,$$

such that if we write $v = (u, j_1, \dots, j_\ell)$, we mean that $v \in (T, S)_{u\downarrow}$, and that v is the j_ℓ-th child of $(u, j_1, \dots, j_{\ell-1})$. Let also

$$s(v) = \sum_{i=1}^{\ell} (\ell - i + 1) j_i,$$

which we denote by s when the node v is understood. Let u_* be the node of S minimizing its value of φ in T, so

$$\varphi(u_*) = \min_{u \in S} \varphi(u).$$

Let K and K' be related such that $K = |\{v \in (T, S)_{u_*\downarrow} \colon s(v) \leq 3K'\}|$. Then,

$$\begin{aligned}
\mathbf{P}\{u_* \notin H_{\varphi;K}(T^\circ)\} &\leq \mathbf{P}\{\exists v \colon s(v) > 3K' \text{ and } \varphi(v) \leq \varphi(u_*)\} \\
&= \mathbf{P}\{\exists u \in S,\, v \in (T, S)_{u\downarrow} \colon s(v) > 3K' \text{ and } \varphi(v) \leq \varphi(u_*)\} \\
&\leq \mathbf{P}\{\exists u \in S,\, v \in (T, S)_{u\downarrow} \colon s(v) > 3K' \text{ and } \varphi(v) \leq \varphi(u)\} \\
&\leq \sum_{u \in V(S)} \mathbf{P}\{\exists v \in (T, S)_{u\downarrow} \colon s(v) > 3K' \text{ and } \varphi(v) \leq \varphi(u)\}, \qquad (7)
\end{aligned}$$

By the arguments of [7, Page 9, Equation (9)], we have

$$\begin{aligned}
&\mathbf{P}\{\exists v \in (T, S)_{u\downarrow} \colon s(v) > 3K' \text{ and } \varphi(v) \leq \varphi(u)\} \\
&\leq 2 \sum_{v \in (T,S)_{u\downarrow} \colon s(v) \in (K', 3K']} \mathbf{P}\{\varphi(v) \leq \varphi(u)\}. \qquad (8)
\end{aligned}$$

Observe now that for $v = (u, j_1, \dots, j_\ell)$,

$$\varphi(v) \leq \varphi(u) \iff \prod_{i=1}^{\ell} |(T, S)_{(u, j_1, \dots, j_i)\downarrow}| \geq \prod_{i=1}^{\ell} (n - |(T, S)_{(u, j_1, \dots, j_i)\downarrow}|).$$

Observe that

$$\frac{|(T, S)_{(u, j_1, \dots, j_i)\downarrow}|}{n} \xrightarrow{d} B \prod_{m=1}^{i} U_{j_m, m}, \qquad (9)$$

as $n \to \infty$, where each $U_{j_m,m}$ is an independent product of j_m independent standard uniform random variables, and $B \sim \mathrm{Beta}(1, k-1)$. So, after dividing through by n,

$$\begin{aligned}
&\mathbf{P}\{\varphi(v) \le \varphi(u)\} \qquad (10)\\
&\quad \le \mathbf{P}\left\{\prod_{i=1}^{\ell} B \prod_{m=1}^{i} U_{j_m,m} \ge \prod_{i=1}^{\ell}\left(1 - B\prod_{m=1}^{i} U_{j_m,m}\right)\right\}\\
&\quad = \mathbf{P}\left\{\prod_{i=1}^{\ell}\prod_{m=1}^{i} U_{j_m,m} \ge \prod_{i=1}^{\ell}\left(\frac{1}{B} - \prod_{i=1}^{i} U_{j_m,m}\right)\right\}\\
&\quad \le \mathbf{P}\left\{\prod_{i=1}^{\ell}\prod_{m=1}^{i} U_{j_m,m} \ge t\right\} + \mathbf{P}\left\{\prod_{i=1}^{\ell}\left(\frac{1}{B} - \prod_{m=1}^{i} U_{j_m,m}\right) \le t\right\}, \qquad (11)
\end{aligned}$$

where $t > 0$ is to be specified later. The first inequality above follows by the portmanteau lemma and the convergence in distribution noted in (9).

In [7, Lemma 1], it is shown that

$$\mathbf{P}\left\{\prod_{i=1}^{\ell}\prod_{m=1}^{i} U_{j_m,m} \ge t\right\} \le \exp\left\{-\sqrt{s/2}\log\left(\frac{s}{e\log(1/t)}\right)\right\}.$$

On the other hand, observing that $1 - e^{-x} \ge (1/2)\min\{x, 1\}$ for all $x \ge 0$, we have

$$\begin{aligned}
&\mathbf{P}\left\{\prod_{i=1}^{\ell}\left(\frac{1}{B} - \prod_{m=1}^{i} U_{j_m,m}\right) \le t\right\}\\
&\quad \le \mathbf{P}\left\{\prod_{i=1}^{\ell}\left(\frac{1}{B} - 1 + \frac{1}{2}\min\left\{\sum_{m=1}^{i}\log(1/U_{j_m,m}), 1\right\}\right) \le t\right\}\\
&\quad \le \mathbf{P}\left\{\prod_{i=1}^{\ell}\left(\frac{1}{B} - 1 + \frac{1}{2}\min\left\{\sum_{m=1}^{i} E_m, 1\right\}\right) \le t\right\}, \qquad (12)
\end{aligned}$$

where $E_1, E_2, \dots$ are independent standard exponential random variables. By the inequality of arithmetic and geometric means,

$$\begin{aligned}
(12) &\le \mathbf{P}\left\{2^{\ell/2}\left(\frac{1}{B} - 1\right)^{\ell/2}\left(\prod_{i=1}^{\ell}\min\left\{\sum_{m=1}^{i} E_m, 1\right\}\right)^{1/2} \le t\right\}\\
&\le \mathbf{P}\left\{2^{\ell}\left(\frac{1}{B} - 1\right)^{\ell} X \le t^2\right\}, \qquad (13)
\end{aligned}$$

where X is defined by

$$X = \prod_{i=1}^{\infty}\min\left\{\sum_{m=1}^{i} E_m, 1\right\}.$$

Then, for $q > 1$ to be specified,

$$(13) \le \mathbf{P}\left\{\frac{B}{1-B} \ge \frac{2}{q^{1/\ell}}\right\} + \mathbf{P}\left\{X \le \frac{t^2}{q}\right\}.$$

By [7, Lemma 2], we know that

$$\mathbf{P}\left\{X \le \frac{t^2}{q}\right\} \le \frac{6t^{1/2}}{q^{1/4}}, \qquad (14)$$

and it is also known that

$$\mathbf{P}\left\{\frac{B}{1-B} \geq \frac{2}{q^{1/\ell}}\right\} = \frac{1}{(1+2/q^{1/\ell})^{k-1}} \leq \exp\left\{-\frac{k-1}{q^{1/\ell}}\right\}, \tag{15}$$

where the inequality follows since $q > 1$, where here we have used that $\frac{1}{1+x} \leq e^{-x/2}$ for $0 \leq x \leq 2$. Optimizing a choice of q in (14) against (15), one can see that when $t > (1/36)e^{-2(k-1)}$,

$$\begin{aligned}(13) &\leq 12t^{1/2}\left(\frac{\log(1/6t^{1/2}) + (\ell/4)\log\left(\frac{k-1}{\log(1/6t^{1/2})}\right)}{k-1}\right)^{\ell/4} \\ &\leq 12t^{1/2}\left(\frac{\log(1/6t^{1/2}) + (\sqrt{2s}/4)\log\left(\frac{k-1}{\log(1/6t^{1/2})}\right)}{k-1}\right)^{\sqrt{2s}/4},\end{aligned}$$

where we used the fact that

$$s = \sum_{i=1}^{\ell}(\ell - i + 1)j_i \geq \ell^2/2.$$

It remains to make an optimal of choice of t. If we pick t such that

$$\log(1/6t^{1/2}) = s^{0.6} - (\sqrt{2s}/4)\log(k-1),$$

it can be shown that if $s > 10^{12}$, $t < 1/6^6$, and $\log(1/6t^{1/2}) > s^{0.6}/2$, then

$$(10) \leq 2\exp\Big\{-\sqrt{s/2}\log\big((1/25)s^{0.3}\log(k-1)\big)\Big\}.$$

Recall the union bound (8),

$$\mathbf{P}\{u_* \notin H_{\varphi;K}(T^\circ)\} \leq 2\sum_{u\in V(S)}\left(\sum_{v\in(T,S)_{u\downarrow}\colon\, s(v)\in(K',3K']}\mathbf{P}\{\varphi(v) \leq \varphi(u)\}\right).$$

We know that for any $u \in V(S)$,

$$|\{v \in (T,S)_{u\downarrow}\colon s(v) \in (K', 3K']\}| \leq 3K'\exp\{\pi\sqrt{2K'}\},$$

(see [7, Page 8, Equation (6)]) and by the conditions imposed on s and k,

$$6kK'\exp\{\pi\sqrt{2K'}\} \leq 6\exp\{11\sqrt{K'}\},$$

and therefore,

$$\begin{aligned}\mathbf{P}\{u_* \notin H_{\varphi;K}(T^\circ)\} &\leq 6\exp\Big\{11\sqrt{K'} - \sqrt{K'/2}\log\big((1/25)K'^{0.3}\log(k-1)\big)\Big\} \\ &\leq 6\exp\Big\{-(1/2)\sqrt{K'}\log\big((1/25)K'^{0.3}\log(k-1)\big)\Big\}\end{aligned}$$

for $K' > 10^{77}$, which holds for $k > 10^{10^8}$. In order to make this probability at most ε, we can pick for some universal constant $c_1 > 0$,

$$\sqrt{K'} = c_1\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon) + \log\log k}.$$

In order to satisfy the condition that $\log(1/6t^{1/2}) > s^{0.6}/2$, this requires that for some constant $c_2 > 0$,

$$\varepsilon < \exp\big\{-c_2(\log k)^{11}\big\}.$$

In this case, there are universal constants $c_3, c_4 > 0$ such that

$$\begin{aligned} K &= |\{v \in (T,S)_{u_*\downarrow} \colon s(v) \leq 3K'\} \\ &\leq c_3 \exp\left\{ c_4 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon) + \log\log k} \right\}. \end{aligned} \qquad \square$$

2.4. **A lower bound.** To obtain a lower bound on $K^*(k, \ell, \varepsilon)$, one can use the likelihood of an observation under $\mathrm{UA}(n,S)$, computed in Section 2.2, and as in [7, Theorem 4], one can then construct a family of probable trees whose maximum likelihood estimate for K-sized sets to intersect the seed will avoid every node of the seed—the right choice for K so that such a tree appears with probability at least ε would then give a lower bound on $K^*(k, \ell, \varepsilon)$. Instead of this lengthy retelling, we use Lemma 2.2 to show how any lower bound on $K(\varepsilon)$ also offers a lower bound on $K^*(k, \ell, \varepsilon)$.

Define $H^*_{K,k}(T^\circ)$ to be the maximum likelihood estimate for the set of size K most likely to contain at least one node of the true seed of $T \sim \mathrm{UA}(n,S)$, given $|S| = k$ and $|L(S)| = \ell$:

$$H^*_{K,k,\ell}(T^\circ) = \operatorname*{argmax}_{H^* \in V(T)^{(K)}} \sum_{S' \in \mathcal{S}_{k,\ell}(T) \colon |V(S') \cap H^*| \geq 1} \mathcal{L}_{T^\circ}(S').$$

In order to prove that $K^*(k, \ell, \varepsilon) \geq K$ for some particular K, it suffices to show that for some specific n, and for all S with $|S| = k$ and $|L(S)| = \ell$,

$$\mathbf{P}_{T \sim \mathrm{UA}(n,S)}\{V(S) \cap H^*_{K,k,\ell}(T^\circ) = \emptyset\} \geq \varepsilon.$$

Proof of Theorem 1.3. Let $m = k^2 K$. For brevity, let $M_u = |(T,S)_{u\downarrow}|$, and let $M_* = \min_{u \in S} M_u$. By Lemma A.1,

$$\frac{M_*}{n} \xrightarrow{d} \frac{\mathrm{Beta}(1, k-1)}{k}$$

as $n \to \infty$. Then,

$$\begin{aligned} \mathbf{P}_{T \sim \mathrm{UA}(m,S)}\{M_*/m > 1/k^2\} &\geq \liminf_{n \to \infty} \mathbf{P}_{T \sim \mathrm{UA}(n,S)}\{M_*/n > 1/k^2\} \\ &\geq \mathbf{P}\{\mathrm{Beta}(1, k-1) > 1/k\} \\ &= (1 - 1/k)^{k-1} \\ &\geq e^{-1}. \end{aligned}$$

Let

$$\mathcal{M} = \left\{ (m_u \colon u \in S) \colon m_u \in \mathbb{N}, \sum_{u \in S} m_u = m, \min_{u \in S} m_u > \frac{m}{k^2} \right\}.$$

Upon conditioning,

$$
\begin{aligned}
&\mathbf{P}_{T\sim \mathrm{UA}(m,S)}\{V(S)\cap H^*_{K,k,\ell}(T^\circ)=\emptyset\}\\
&\quad\geq \mathbf{P}\{M_*/m > 1/k^2\}\\
&\quad\quad \sum_{(m_u\colon u\in S)\in\mathcal{M}} \mathbf{P}\left\{\bigcap_{u\in S}[u\notin H^*_{K,k,\ell}(T^\circ)\cap (T,S)_{u\downarrow}]\;\middle|\;\bigcap_{u\in S}[M_u=m_u]\right\}\\
&\quad\geq e^{-1}\sum_{(m_u\colon u\in S)\in\mathcal{M}}\left(\prod_{u\in S}\mathbf{P}\Big\{u\notin H^*_{K,k,\ell}((T,S)^\circ_{u\downarrow})\;\Big|\;M_u=m_u\Big\}\right)\\
&\quad\geq e^{-1}\Big(\mathbf{P}_{T\sim\mathrm{UA}(K+1)}\{u_1\notin H^*_{K,1}(T^\circ)\}\Big)^k,
\end{aligned}
$$

where this last line follows from the optimality of $H^*_{K,1}((T,S)^\circ_{u\downarrow})$ as a root estimator in $(T,S)^\circ_{u\downarrow}$, and since conditionally upon $M_u=m_u$, $(T,S)_{u\downarrow}$ is distributed as $\mathrm{UA}(m_u)$ by Lemma 2.2. By definition, if $K<K((e\varepsilon)^{1/k})$, then the probability that $H^*_{K,k,\ell}(T^\circ)$ avoids $V(S)$ exceeds ε. □

Corollary 1.4 follows immediately.

3. Finding all seed vertices

3.1. **Upper bounds.**

3.1.1. *A familiar strategy.* We can also use the strategy $H^*_{\psi;K}$ of [7] and Section 2.1 to get all the nodes of S; in this section, when the procedure is used to find all nodes of the seed, we omit the asterisk for notational consistency. More specifically, we study the smallest choice of K for which $H_{\psi;K}$ contains all nodes of S with probability at least $1-\varepsilon$, and such a choice will give an upper bound on $K(k,\ell,\varepsilon)$.

If $\psi(u)=|(T,u)_{v\downarrow}|$ for v adjacent to u, we will say that $\psi(u)$ is *witnessed at* v.

Proposition 3.1. *There are universal constants $c,\varepsilon_0>0$ such that, if $\varepsilon\leq\varepsilon_0$ and*

$$K\geq (ck\ell/\varepsilon)\log(k\ell/\varepsilon),$$

then

$$\mathbf{P}\{V(S)\subseteq H_{\psi;K}(T^\circ)\}\geq 1-\varepsilon.$$

Proof. The proof is similar to that of Proposition 2.1. Write

$$\psi^*=\max\{\psi(u_1),\ldots,\psi(u_k)\}.$$

Observe that if for all $i>K$, $\psi(u_i)>\psi^*$, then $V(S)\subseteq H_{\psi;K}(T^\circ)$. So,

$$\mathbf{P}\{V(S)\not\subseteq H_{\psi;K}(T^\circ)\}\leq \mathbf{P}\{\psi^*\geq nt\}+\mathbf{P}\{\exists i>K\colon \psi(u_i)\leq nt\} \tag{16}$$

for $t>0$ to be specified. We handle the first term in (16): Suppose that ψ^* is attained by $u\in V(S)$ and witnessed by its child $v\notin V(S)$, *i.e.,*

$$\psi^*=\psi(u)=|(T,u)_{v\downarrow}|,$$

Then, u has a neighbour $w\in V(S)$, and

$$\psi(w)\geq |(T,w)_{u\downarrow}|>|(T,u)_{v\downarrow}|=\psi(u),$$

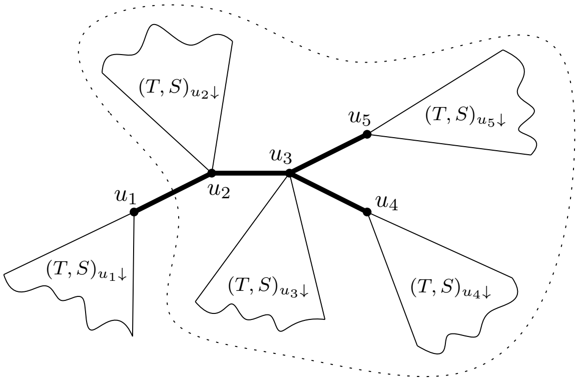


FIGURE 3. The seed S has vertex set $\{u_1, \ldots, u_5\}$. Suppose that $\psi^* = \psi(u_1)$. Then, ψ^* is witnessed at u_2, and ψ^* correponds to the size of the outlined subgraph.

so $\psi(u)$ cannot be maximum. Thus, ψ^* must be witnessed by a node of $V(S)$, and in particular, one can iterate the above motion to see that ψ^* must be attained by a leaf of S and witnessed by its unique neighbour in S, so

$$\psi^* = \max_{u \in L(S)} \sum_{v \in V(S)-u} |(T,S)_{v\downarrow}|.$$

See Figure 3 for an illustration. By Lemma A.2, for any $u \in L(S)$,

$$\frac{1}{n} \sum_{v \in V(S)-u} |(T,S)_{v\downarrow}| \xrightarrow{d} \text{Beta}(k-1,1),$$

as $n \to \infty$, so

$$\begin{aligned} \mathbf{P}\{\psi^* \geq nt\} &\leq \mathbf{P}\left\{ \exists u \in L(S) \colon \sum_{v \in V(S)-u} |(T,S)_{v\downarrow}| \geq nt \right\} \\ &\leq \ell\, \mathbf{P}\{\text{Beta}(k-1,1) \geq t\} \\ &= \ell(1-t^{k-1}). \end{aligned}$$

We can make this at most $\varepsilon/2$ by choosing $t = (1-\varepsilon/(2\ell))^{1/(k-1)}$.

For the second term in (16), the argument is again identical to that of [7, Theorem 3], and we can say that

$$\mathbf{P}\{\exists i > K \colon \psi(u_i) \leq nt\} \leq Kt^{K-1} \leq Ke^{-\frac{\varepsilon(K-1)}{2(k-1)\ell}}.$$

Picking $K \geq (ck\ell/\varepsilon)\log(k\ell/\varepsilon)$ for some constant $c > 0$ gives the desired result, as long as $\varepsilon \leq \varepsilon_0$ for some constant $\varepsilon_0 > 0$. □

Again, the set $H_{\psi;K}(T^\circ)$ can be computed in $O(n)$ time. We show in Lemma 4.2 that the result of Theorem 1.5 involves the right order for K for the strategy given by $H_{\psi;K}$, up to logarithmic factors: When $K \leq k\ell/(4\varepsilon)$, then with probability at least ε, at least one leaf of S is also a leaf of $T \sim \text{UA}(K,S)$, and any leaf of T maximizes the value of ψ.

3.1.2. *A reduction to intersection testing.* We now consider an alternative procedure for locating all nodes of the seed, which sometimes requires fewer nodes than $H_{\psi;K}$ to succeed with probability $1-\varepsilon$. We define the set $H_{\phi;K,k,\ell,\varepsilon}(T^\circ)$ as follows:

(i) Let H^* be a set of size $K^*(k,\ell,\varepsilon/2)$ which intersects the seed with probability at least $1-\varepsilon/2$;
(ii) For each $u \in H^*$, traverse the tree T in a depth-first manner around u;
(iii) When exploring $v \in T$, add v to $H_{\phi;K,k,\ell,\varepsilon}(T^\circ)$ if $|(T,u)_{v\downarrow}| \geq n\varepsilon/(2k\ell)$, and stop exploring this path otherwise. Stop at any point if the size of the set exceeds K.

If we can prove that $H_{\phi;K,k,\ell,\varepsilon}(T^\circ)$ includes the whole seed with probability at least $1-\varepsilon$, and that K is sufficiently large, we obtain an upper bound on $K(k,\ell,\varepsilon)$.

Proposition 3.2. *If $K \geq (2k\ell/\varepsilon)K^*(k,\ell,\varepsilon/2)$, then*

$$\mathbf{P}\{V(S) \subseteq H_{\phi;K,k,\ell,\varepsilon}(T^\circ)\} \geq 1-\varepsilon.$$

Proof. We show the failure probability is small enough:

$$\mathbf{P}\{V(S) \not\subseteq H_{\phi;K,k,\ell,\varepsilon}(T^\circ)\} \leq \mathbf{P}\bigg\{V(S) \cap H^* = \emptyset\bigg\} + \mathbf{P}\bigg\{\min_{u\in L(S)} |(T,S)_{u\downarrow}| < \frac{n\varepsilon}{2k\ell}\bigg\}.$$

The first term is at most $\varepsilon/2$ by definition. For the second term, note that

$$\frac{\min_{u\in L(S)} |(T,S)_{u\downarrow}|}{n} \xrightarrow{d} \frac{\mathrm{Beta}(1,k-1)}{\ell},$$

as $n \to \infty$ by Lemma A.1, so

$$\begin{aligned}\mathbf{P}\bigg\{\min_{u\in L(S)} |(T,S)_{u\downarrow}| < \frac{n\varepsilon}{2k\ell}\bigg\} &\leq \mathbf{P}\bigg\{\mathrm{Beta}(1,k-1) \leq \frac{\varepsilon}{2k}\bigg\} \\ &= 1-\left(1-\frac{\varepsilon/2}{k}\right)^{k-1} \\ &\leq 1-e^{-\varepsilon/2} \\ &\leq \varepsilon/2,\end{aligned}$$

as desired. Finally, for u fixed, there are at most $2k\ell/\varepsilon$ nodes v such that $|(T,u)_{v\downarrow}| \geq n\varepsilon/(2k\ell)$, so K is indeed large enough to include all desired nodes. □

By a remark in Section 2.2, the set H^* in the above construction can be computed in polynomial time, so $H_{\phi;K,k,\ell,\varepsilon}(T^\circ)$ can be computed in polynomial time. Theorem 1.5 follows immediately from Proposition 3.1 and Proposition 3.2.

3.2. **Lower bounds.** By (1), we have the same lower bound on $K(k,\ell,\varepsilon)$ as in Theorem 1.3, *i.e.,*

Corollary 3.3. *There are universal constants $c_1, c_2, c_3 > 0$, and such that if $\varepsilon \leq e^{-c_1 k}$, then*

$$K(k,\ell,\varepsilon) \geq c_2 \exp\bigg\{c_3\sqrt{\frac{\log(1/\varepsilon)}{k}}\bigg\}.$$

Unlike the case for $K^*(k,\ell,\varepsilon)$, we do not know that $K(k,\ell,\varepsilon)$ is at most linear in k when $\varepsilon > e^{-ck}$; our best upper bound from Theorem 1.5 says that $K(k,\ell,\varepsilon)$ is at most exponential in k, while the lower bound from Corollary 3.3 does not apply. We thus search for a lower bound on $K(k,\ell,\varepsilon)$ which applies in the regime when ε is large.

As in Section 2.4, define $H_{K,k,\ell}(T^\circ)$ to be the set of size K most likely to contain all the nodes of the true seed of $T \sim \mathrm{UA}(n,S)$, given $|S|=k$ and $|L(S)|=\ell$:

$$H_{K,k,\ell}(T^\circ) = \operatorname*{argmax}_{H \in V(T)^{(K)}} \sum_{S' \in \mathcal{S}_{k,\ell}(T)\colon V(S') \subseteq H} \mathcal{L}_{T^\circ}(S').$$

If ε is at least some positive constant, then $K(k,\ell,\varepsilon) \le ck\ell$ for some constant $c > 0$ by Theorem 1.5. We noted in Section 3.1.1, as a result of Lemma 4.2, that if $H_{\psi;K}$ were the optimal strategy for picking K nodes to include the whole seed with probability at least $1-\varepsilon$, then $k\ell/(4\varepsilon)$ is roughly a lower bound on the size of such a vertex confidence set. We know that $H_{\psi;K}$ is not in fact the optimal strategy, but one can understand it to be a relaxation of the optimal strategy $H_{K,k,\ell}$. We thus make the following conjecture, which expresses our belief that $H_{\psi;K}$ is "close enough" to $H_{K,k,\ell}$ when ε is large.

Conjecture 3.4. *There are universal constants $c_1, c_2, \varepsilon_0 > 0$ such that if $k, \ell > c_1$, then,*

$$K(k,\ell,\varepsilon_0) \ge c_2 k\ell.$$

4. Finding all leaves given the skeleton

For a seed S, write $R(S) = S - L(S)$ for the *skeleton* of S, or simply R when S is understood. For an integer $i \ge 1$, let $\mathcal{T}_i$ denote the set of trees in which a set of i labelled vertices form a connected subgraph, and in which all other vertices are unlabelled. For a given labelled tree T and a set $A \in V(T)^{(i)}$, let $T^{(A)} \in \mathcal{T}_i$ be the tree in which all labels are forgotten except for those of A. Consider now the optimal size

$$K'(k,\ell,\varepsilon) = \min\left\{ m\colon \begin{array}{l} \exists H'_{m,k,\ell,\varepsilon}\colon \mathcal{T}_{k-\ell} \to V(\mathcal{T}_{k-\ell})^{(m)}, \text{ such that} \\ \displaystyle\min_{\substack{S\colon |S|=k \\ |L(S)|=\ell}} \mathbf{P}_{T\sim\mathrm{UA}(n,S)}\{L(S) \subseteq H'_{m,k,\ell,\varepsilon}(T^{(R)})\} \ge 1-\varepsilon \end{array} \right\}.$$

be the optimal size of a set which, given the position of the skeleton of the seed, the size of the seed, and its number of leaves, will locate all of its true leaves with probability at least $1-\varepsilon$.

As in Section 2 and Section 3, we find an upper bound on $K'(k,\ell,\varepsilon)$ by exhibiting an algorithm which, given k, ε and R, returns a set of vertices which contains all of $L(S)$ with probability at least $1-\varepsilon$.

Let $\psi(u) = |(T,R)_{u\downarrow}|$, and let $H'_{\psi;K}(T^{(R)})$ be the set of K vertices $u \in N(R)$ maximizing their value of ψ. This estimator for $L(S)$ is slightly different than those for $V(S)$ seen in Section 2 and Section 3. Indeed, allowing ourselves to assume R significantly improves our chances at correctly guessing the rest of S. Specifically, if k is constant, the dependence of $K'(k,\ell,\varepsilon)$ upon $1/\varepsilon$ is shown to be logarithmic, while the result of Theorem 1.3 has that, for sufficiently small ε, $K^*(k,\ell,\varepsilon)$ is superpolylogarithmic in $1/\varepsilon$.

Proposition 4.1. *If*

$$K \ge \ell + 2(k-\ell)\log\left((3\ell/\varepsilon)\log\left(\frac{3(k-\ell)}{\varepsilon}\right)\right) + (7/6)\log\left(\frac{3(k-\ell)}{\varepsilon}\right)$$

then $\mathbf{P}\{L(S) \subseteq H'_{\psi;K}(T^{(R)})\} \ge 1-\varepsilon$.

Proof. Let $v_1, v_2, \dots$ be the chronological sequence of nodes attaching to R, where $\{v_1, v_2, \dots, v_\ell\} = L(S)$ ordered arbitrarily. Write $\psi_* = \min_{u \in L(S)} \psi(u)$. If for all $i > K$, $\psi(v_i) < \psi_*$, then $L(S) \subseteq H'_{\psi;K}(T^{(R)})$. So

$$\mathbf{P}\{L(S) \not\subseteq H'_{\psi;K}(T^{(R)})\} \le \mathbf{P}\{\psi_* \le tn\} + \mathbf{P}\{\exists i > K \colon \psi(v_i) \ge tn\} \tag{17}$$

for $t > 0$ to be specified. Observe that ψ_*/n converges in distribution to $\mathrm{Beta}(1, k-1)/\ell$ as $n \to \infty$, so

$$\begin{aligned}\mathbf{P}\{\psi_* \le tn\} &\le \mathbf{P}\{\mathrm{Beta}(1, k-1) \le t\ell\}\\ &\le t\ell(k-1)\\ &\le \varepsilon/3\end{aligned}$$

if we choose $t = \varepsilon/(3\ell(k-1))$.

It remains to handle the second term in (17). Let N be the (random) time at which v_K is inserted, *i.e.,* $v_K = u_N$. Let T_u be the component of T containing u after the removal of all edges between $\{u_1, \dots, u_N\}$. Any node v_i with $i > K$ is part of T_u for some $u \in R$. Since for any $u \in R$, $|T_u|/n$ converges in distribution to $\mathrm{Beta}(1, N-1)$ as $n \to \infty$,

$$\begin{aligned}&\mathbf{P}\{\exists i > K \colon \psi(v_i) \ge tn\}\\ &\quad\le \mathbf{P}\{\exists u \in R \colon |T_u| \ge nt\}\\ &\quad\le (k-\ell)\,\mathbf{P}\{\mathrm{Beta}(1, N-1) \ge t\}\\ &\quad\le (k-\ell)\,\mathbf{P}\{\mathrm{Beta}(1, N-1) \ge t \mid N \ge s\} + (k-\ell)\,\mathbf{P}\{N \le s\}, \qquad (18)\end{aligned}$$

where $s > 0$ is to be specified. Conditionally upon $N \ge s$, $\mathrm{Beta}(1, s-1)$ stochastically dominates $\mathrm{Beta}(1, N-1)$ by Lemma A.4, so

$$\begin{aligned}(k-\ell)\,\mathbf{P}\{\mathrm{Beta}(1, N-1) \ge t \mid N \ge s\} &\le (k-\ell)\,\mathbf{P}\{\mathrm{Beta}(1, s-1) \ge t\}\\ &= (k-\ell)(1-t)^{s-1}\\ &\le (k-\ell)e^{-t(s-1)},\end{aligned}$$

which is less than $\varepsilon/3$ if we choose

$$s = 1 + (1/t)\log\left(\frac{3(k-\ell)}{\varepsilon}\right) = 1 + \frac{3\ell(k-1)}{\varepsilon}\log\left(\frac{3(k-\ell)}{\varepsilon}\right).$$

For the second term in (18), observe that

$$\mathbf{P}_{T\sim\mathrm{UA}(n,S)}\{N \le s\} \le \mathbf{P}_{T\sim\mathrm{UA}(s,S)}\{\deg(R) \ge K\},$$

where

$$\deg(R) = \ell + \sum_{i=k+1}^{s} \mathbf{1}\{u_i \text{ connects to } R\}$$

and each such indicator is independent. Clearly, for any $i \ge k+1$,

$$\mathbf{1}\{u_i \text{ connects to } R\} \sim \mathrm{Bernoulli}\left(\frac{k-\ell}{i-1}\right),$$

so writing $H_m = \sum_{i=1}^m 1/i$ for the m-th Harmonic number,

$$\mathbf{E}\{\deg(R)\} = \ell + (k-\ell)(H_{s-1} - H_{k-1}).$$

Picking

$$K \ge \ell + (k-\ell)(H_{s-1} - H_{k-1}) + \delta$$

for $\delta > 0$ to be specified, we have by Bernstein's inequality [3, 6],

$$\begin{aligned}
&\mathbf{P}\{\deg(R) \geq K\} \\
&\quad \leq \mathbf{P}\left\{ \left[\sum_{i=k+1}^{s} \text{Bernoulli}\left(\frac{k-\ell}{i-1} \right) \right] - (k-\ell)(H_{s-1} - H_{k-1}) \geq \delta \right\} \\
&\quad \leq \exp\left\{ - \frac{\delta^2}{2(k-\ell)(H_{s-1} - H_{k-1}) + 2\delta/3} \right\} \\
&\quad \leq \exp\left\{ - \frac{\delta^2}{2(k-\ell)\log\left(\frac{s-1}{k}\right) + 2\delta/3} \right\}.
\end{aligned}$$

Some arithmetic shows that picking

$$\delta = (2/3)\log\left(\frac{3(k-\ell)}{\varepsilon}\right) + \sqrt{2(k-\ell)\log\left(\frac{s-1}{k}\right)\log\left(\frac{3(k-\ell)}{\varepsilon}\right)}$$

suffices to have

$$\mathbf{P}\{\deg(R) \geq K\} \leq \frac{\varepsilon}{3(k-\ell)},$$

and therefore, in (18),

$$(k-\ell)\,\mathbf{P}\{N \leq s\} \leq \varepsilon/3.$$

With our particular choice of s, this proves that it suffices to pick

$$K \geq \ell + 2(k-\ell)\log\left((3\ell/\varepsilon)\log\left(\frac{3(k-\ell)}{\varepsilon}\right)\right) + (7/6)\log\left(\frac{3(k-\ell)}{\varepsilon}\right). \qquad \square$$

Theorem 1.6 follows immediately. In fact, Proposition 4.1 is tight for a large class of seeds. In order to prove this, we use the following basic result.

Lemma 4.2. *When $K \leq k\ell/(4\varepsilon)$,*

$$\mathbf{P}_{T\sim \mathrm{UA}(K,S)}\{\exists u \in L(S) \colon |(T,S)_{u\downarrow}| = 1\} \geq \varepsilon.$$

Proof. Since S is a tree, it has at least two leaves. Write $\mathcal{E}_u$ for the event that $|(T,S)_{u\downarrow}| = 1$. By inclusion-exclusion, for some arbitrary distinct leaves $u, v \in L(S)$,

$$\mathbf{P}\{\exists w \in L(S) \colon \mathcal{E}_w\} \geq \ell\,\mathbf{P}\{\mathcal{E}_u\} - \binom{\ell}{2}\mathbf{P}\{\mathcal{E}_u \cap \mathcal{E}_v\}.$$

It is easy to see that

$$\mathbf{P}\{\mathcal{E}_u\} = \frac{k-1}{k} \cdot \frac{k}{k+1} \cdots \frac{K-2}{K-1} = \frac{k-1}{K-1},$$

and

$$\mathbf{P}\{\mathcal{E}_u \cap \mathcal{E}_v\} = \frac{k-2}{k} \cdot \frac{k-1}{k+1} \cdots \frac{K-3}{K-1} = \frac{(k-2)(k-1)}{(K-2)(K-1)} \leq \left(\frac{k-1}{K-1}\right)^2,$$

so

$$\mathbf{P}\{\exists w \in L(S) \colon \mathcal{E}_w\} \geq \frac{k\ell}{4K}. \qquad \square$$

Once again, we rely on the maximum likelihood estimate to prove a lower bound. Let $H'_{K,k,\ell}(T^{(R)})$ be the maximum likelihood estimate for K-sized sets to include all leaves of a seed S, given the skeleton R, and given $|S| = k$ and $|L(S)| = \ell$, *i.e.*,

$$H'_{K,k,\ell}(T^{(R)}) = \operatorname*{argmax}_{H \in N(R)^{(K)}} \sum_{L' \subseteq H \colon |L'| = \ell} \mathcal{L}_{T^{(R)}}(L'),$$

where $\mathcal{L}_{T^{(R)}}(L')$ represents the likelihood of observing the tree $T^{(R)}$ if it were drawn from $\mathrm{UA}(n, R \cup L')^{(R)}$.

Proposition 4.3. *Let $\ell_2 = |L(R)|$. Suppose that $k - \ell - \ell_2 \geq 2\sqrt{k}$, $\ell_2 \geq 2\sqrt{k}$, and*

$$\varepsilon \leq \frac{\ell}{128 e^5 \ell_2^4}.$$

Then, there is a universal constant $c > 0$ such that if

$$K \leq c(k - \ell - \ell_2) \log(\ell_2 \ell / \varepsilon),$$

then

$$\mathbf{P}_{T \sim \mathrm{UA}(n,S)}\{L(S) \subseteq H'_{K,k,\ell}(T^{(R)})\} < 1 - \varepsilon.$$

Proof. Let $\mathcal{E}_u$ be the event that $|(T,S)_{u\downarrow}| = 1$. By Lemma 4.2, when $n = k\ell/(64\varepsilon)$,

$$\mathbf{P}_{T \sim \mathrm{UA}(n,S)}\{\exists u \in L(S) \colon \mathcal{E}_u\} \geq 16\varepsilon.$$

Let

$$X = |\{v \in N(R - L(R)) - S \colon |(T,S)_{u\downarrow}| \geq 2\}|.$$

For $u \in L(R)$, let $\mathcal{F}_u$ be the event that $|N(u) - S| \geq 1$, and let $\mathcal{F} = \cap_{u \in L(R)} \mathcal{F}_u$. Then,

$$\begin{aligned}
&\mathbf{P}\{L(S) \not\subseteq H'_{K,k,\ell}(T^{(R)})\} \\
&\qquad \geq 16\varepsilon\, \mathbf{P}\big\{L(S) \not\subseteq H_{K,k,\ell}(T^{(R)}) \bigm| \exists u \in L(S) \colon \mathcal{E}_u\big\} \\
&\qquad \geq 16\varepsilon\, \mathbf{P}\big\{[X \geq K] \cap \mathcal{F} \bigm| \exists u \in L(S) \colon \mathcal{E}_u\big\} \\
&\qquad \geq 16\varepsilon\, \mathbf{P}\big\{[X \geq K] \cap \mathcal{F}\big\},
\end{aligned}$$

where the second inequality follows since the X vertices u of $N(R - L(R)) - S$ with $|(T,S)_{u\downarrow}| \geq 2$ are more likely than at least one vertex of $L(S)$ with $|(T,S)_{u\downarrow}| = 1$ of being chosen in $H'_{K,k,\ell}(T^{(R)})$ when $\mathcal{F}$ holds, and the third inequality follows since conditioning on the seed's leaves to be naked can only increase the likelihood of connections to skeleton nodes. It suffices to prove that $\mathbf{P}\{[X \geq K] \cap \mathcal{F}\} \geq 1/16$.

Let

$$T_R = \bigcup_{u \in R - L(R)} (T,S)_{u\downarrow}, \quad T_L = \bigcup_{u \in L(R)} (T,S)_{u\downarrow},$$

with sizes $|T_R| = M_R$ and $|T_L| = M_L$. Observe that, conditionally upon the sizes M_R and M_L, the events $X \geq K$ and $\mathcal{F}$ are independent. Furthermore,

$$\frac{M_R}{n} \xrightarrow{d} B_R \sim \mathrm{Beta}(k - \ell - \ell_2, \ell + \ell_2), \quad \frac{M_L}{n} \xrightarrow{d} B_L \sim \mathrm{Beta}(\ell_2, k - \ell_2),$$

as $n \to \infty$, where we note that B_R and B_L are not necessarily independent. Let $\mathcal{M} \subseteq \mathbb{N}^2$ be such that for all $(m_R, m_L) \in \mathcal{M}$,

$$\left| \frac{m_R}{n} - \frac{k - \ell - \ell_2}{k} \right| < \frac{1}{\sqrt{k}} \quad \text{and} \quad \left| \frac{m_L}{n} - \frac{\ell_2}{k} \right| < \frac{1}{\sqrt{k}}.$$

Write $M = (M_R, M_L)$ for brevity. By Lemma A.5,

$$\begin{aligned}
\mathbf{P}\{M \in \mathcal{M}\} &= \mathbf{P}\left\{\left|\frac{M_R}{n} - \frac{k-\ell-\ell_2}{k}\right| < \frac{1}{\sqrt{k}}, \left|\frac{M_L}{n} - \frac{\ell_2}{k}\right| < \frac{1}{\sqrt{k}}\right\} \\
&\geq \mathbf{P}\left\{\left|B_R - \frac{k-\ell-\ell_2}{k}\right| < \frac{1}{\sqrt{k}}, \left|B_L - \frac{\ell_2}{k}\right| < \frac{1}{\sqrt{k}}\right\} \\
&\geq 1 - \mathbf{P}\left\{\left|B_R - \frac{k-\ell-\ell_2}{k}\right| \geq \frac{1}{\sqrt{k}}\right\} - \mathbf{P}\left\{\left|B_L - \frac{\ell_2}{k}\right| \geq \frac{1}{\sqrt{k}}\right\} \\
&\geq 1/2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbf{P}\{[X \geq K] \cap \mathcal{F}\} \\
&\quad\geq \sum_{m \in \mathcal{M}} \mathbf{P}\big\{[X \geq K] \cap \mathcal{F} \bigm| M = m\big\}\, \mathbf{P}\{M = m\} \\
&\quad= \sum_{m \in \mathcal{M}} \mathbf{P}\big\{X \geq K \bigm| M = m\big\}\, \mathbf{P}\big\{\mathcal{F} \bigm| M = m\big\}\, \mathbf{P}\{M = m\} \\
&\quad= \sum_{m \in \mathcal{M}} \mathbf{P}_{T \sim \mathrm{UA}(n,S)}\{M = m\} \\
&\qquad\quad \mathbf{P}_{T \sim \mathrm{UA}(m_R, R - L(R))}\{\deg(R - L(R)) \geq K\} \\
&\qquad\quad \mathbf{P}_{T \sim \mathrm{UA}(m_L, L(R))}\{\forall u \in L(R),\, \deg(u) \geq 1\} \\
&\quad\geq (1/2)\, \mathbf{P}_{T \sim \mathrm{UA}(n(k-\ell-\ell_2-\sqrt{k})/k, R - L(R))}\{\deg(R - L(R)) \geq K\} \\
&\qquad\quad \mathbf{P}_{T \sim \mathrm{UA}(n(\ell_2-\sqrt{k})/k, L(R))}\{\forall u \in L(R),\, \deg(u) \geq 1\}.
\end{aligned}$$

Let $v_1, \ldots, v_{n(k-\ell-\ell_2-\sqrt{k})/k}$ be the sequence of nodes connecting to $R - L(R)$ in the uniform attachment process implied by the first probability above. Let

$$X_i = \mathbf{1}\{v_i \text{ connects to } R \text{ and } v_j \text{ connects to } u_i \text{ for some } j > i\}.$$

Then,

$$\deg(R - L(R)) = \sum_{i=k-\ell-\ell_2+1}^{n(k-\ell-\ell_2-\sqrt{k})/k} X_i,$$

where $\{X_i \colon k-\ell-\ell_2+1 \leq i \leq n(k-\ell-\ell_2-\sqrt{k})/k\}$ is a collection of independent Bernoulli random variables with

$$\mathbf{E}\{X_i\} = (k-\ell-\ell_2)\left(\frac{1}{i-1} - \frac{1}{\frac{n(k-\ell-\ell_2-\sqrt{k})}{k} - 1}\right).$$

Since $k - \ell - \ell_2 \geq 2\sqrt{k}$ by assumption, we can see that

$$\mathbf{E}\{\deg(R - L(R))\} \geq (k-\ell-\ell_2)\log\left(\frac{n}{2e^2k}\right).$$

Since each X_i is independent, then

$$\begin{aligned}\mathbf{Var}\{\deg(R-L(R))\} &= \sum_{i=k-\ell-\ell_2+1}^{n(k-\ell-\ell_2-\sqrt{k})/k} \mathbf{Var}\{X_i\}\\ &\le \sum_{i=k-\ell-\ell_2+1}^{n(k-\ell-\ell_2-\sqrt{k})/k} \mathbf{E}\{X_i\}\\ &= \mathbf{E}\{\deg(R-L(R))\}.\end{aligned}$$

Since the median of $\deg(R-L(R))$ is within a standard deviation of its mean, we see that picking

$$K \le \left(\frac{k-\ell-\ell_2}{2}\right)\log\left(\frac{n}{2e^2k}\right)$$

is sufficient to make

$$\mathbf{P}_{T\sim \mathrm{UA}(n(k-\ell-\ell_2-\sqrt{k})/k, R-L(R))}\{\deg(R-L(R)) \ge K\} \ge 1/2,$$

as long as $(k-\ell-\ell_2)\log\left(\frac{n}{2e^2k}\right) \ge 4$. Since $k-\ell-\ell_2 \ge 2\sqrt{k} \ge 2$, this condition is satisfied as long as $n \ge e^4 k$. This condition will be absorbed by a further condition on n, and can be safely ignored.

Let $w_1, \dots, w_{n(\ell_2-\sqrt{k})/k}$ be the chronological sequence of nodes appearing in the attachment process $T \sim \mathrm{UA}(n(\ell_2-\sqrt{k})/k, L(R))$, and let $z_1, \dots, z_Y$ be the subsequence of these nodes which connect directly to the nodes of $L(R)$. Finally, let Z be minimum such that the nodes $z_1, \dots, z_Z$ connect to all nodes of $L(R)$. Then,

$$\mathbf{P}_{T\sim \mathrm{UA}(n(\ell_2-\sqrt{k})/k, L(R))}\{\forall u \in L(R),\, \deg(u) \ge 1\} \ge \mathbf{P}\{Z \le Y\}.$$

Then, writing

$$Y_i = \mathbf{1}\{w_i \text{ connects to a node of } L(R)\},$$

then we see that $\{Y_i \colon \ell_2+1 \le i \le n(\ell_2-\sqrt{k})/k\}$ is a collection of independent Bernoulli random variables such that $Y_i \sim \mathrm{Bernoulli}\left(\frac{\ell_2}{i-1}\right)$ and

$$Y = \sum_{i=\ell_2+1}^{n(\ell_2-\sqrt{k})/k} Y_i.$$

Just as before, we see that if $\ell_2 \ge 2\sqrt{k}$, then $\mathbf{E}\{Y\} \ge \ell_2 \log\left(\frac{n}{2ek}\right)$, and

$$\mathbf{P}\left\{Y \ge (\ell_2/2)\log\left(\frac{n}{2ek}\right)\right\} \ge 1/2,$$

whence

$$\begin{aligned}\mathbf{P}\{\forall u \in L(R),\, \deg(u) \ge 1\} &\ge (1/2)\,\mathbf{P}\left\{Z \le Y \;\middle|\; Y \ge (\ell_2/2)\log\left(\frac{n}{2ek}\right)\right\}\\ &\ge (1/2)\,\mathbf{P}\left\{Z \le (\ell_2/2)\log\left(\frac{n}{2ek}\right)\right\}.\end{aligned}$$

The random variable Z is distributed as the time to collect all coupons in the well-known *coupon collector* problem [22, Section 2.4.1]. Specifically,

$$Z \sim \sum_{i=1}^{\ell_2} \mathrm{Geo}\left(\frac{\ell_2-(i-1)}{\ell_2}\right),$$

where each term above is independent, and where $\mathrm{Geo}(p)$ denotes a geometric random variable with parameter p, *i.e.,* the random variable with probability mass function $f_p\colon \mathbb{N} \to \mathbb{R}$, where

$$f_p = p(1-p)^{k-1}.$$

Then, by Markov's inequality,

$$\begin{aligned}\mathbf{P}\Big\{Z \le (\ell_2/2)\log\Big(\frac{n}{2ek}\Big)\Big\} &= 1 - \mathbf{P}\Big\{Z > (\ell_2/2)\log\Big(\frac{n}{2ek}\Big)\Big\}\\ &\ge 1 - \frac{\ell_2\log(e\ell_2)}{(\ell_2/2)\log\big(\frac{n}{2ek}\big)}\\ &\ge 1/2\end{aligned}$$

as long as $n \ge 2e^5\ell_2^4 k$. Finally, this proves that

$$\mathbf{P}\{[X > K] \cap \mathcal{F}\} \ge 1/16. \qquad \square$$

Note that the family of complete binary trees satisfies the structural seed conditions of Proposition 4.3. Indeed, if S is a complete binary tree, we have that $k = 2\ell - 1$ and $\ell = 2\ell_2 - 1$, so that

$$k - \ell - \ell_2 = k - \frac{k+1}{2} - \frac{k+3}{4} = \frac{k-5}{4} \ge 2\sqrt{k},$$

and

$$\ell_2 = \frac{k+3}{4} \ge 2\sqrt{k},$$

for all $k \ge 74$.

5. Finding a whole star

In this section, we study the number of nodes $K(S_k, \varepsilon)$ required to find the seed in a seeded uniform attachment tree with probability at least $1-\varepsilon$, given that the seed is a star S_k on k nodes, and given that we know that the seed is isomorphic to S_k. By (3) and Theorem 1.5,

$$K(S_k, \varepsilon) \le ck^2(1/\varepsilon)^{1+2/k}\log(1/\varepsilon).$$

The extra knowledge of the full structure of S_k allows us to shave off an extra factor of k/ε.

Let u_1 denote the center of the star. To identify the whole star, we first locate u_1 and then use Proposition 4.1 to locate all remaining vertices. Recall $H^*_{\psi;m}$ from Section 2.1. We use the following intermediate result, whose proof is adapted from that of Theorem 1.1.

Lemma 5.1. *There are universal constants $c, \varepsilon_0 > 0$ such that if $\varepsilon \le \varepsilon_0$ and*

$$K \ge c(1/\varepsilon)^{1/k}\log(1/\varepsilon),$$

then

$$\mathbf{P}\{u_1 \in H^*_{\psi;K}(T^\circ)\} \ge 1 - \varepsilon.$$

Proof. If for all $i > K$, $\psi(u_i) > \psi(u_1)$, then $H^*_{\psi;K}(T^\circ)$ contains u_1. Moreover, if T_i denotes the component of T containing u_i after the removal of all edges between vertices of S_k,

$$\begin{aligned}\mathbf{P}\{\psi(u_1) \geq nt\} &\leq \mathbf{P}\{\exists 1 \leq i \leq k \colon |T_i| \geq nt\}\\ &\leq k\,\mathbf{P}\{\mathrm{Beta}(1, k-1) \geq t\}\\ &= k(1-t)^{k-1}\end{aligned}$$

and this probability is at most $\varepsilon/2$ for $t = 1 - (\varepsilon/(2k))^{1/(k-1)}$. As before,

$$\mathbf{P}\{\exists i > K \colon \psi(u_i) \leq nt\} \leq Kt^{K-1} \leq Ke^{-(K-1)\left(\frac{\varepsilon}{2k}\right)^{1/(k-1)}},$$

and it is not hard to see that this probability can be made at most $\varepsilon/2$ by choosing, for some constant $c > 0$,

$$K \geq c(1/\varepsilon)^{1/k} \log(1/\varepsilon). \qquad \square$$

We note here a related result by Jog and Loh [18, Theorem 4], which says that for a universal constant $c > 0$, if $k \geq c \log(1/\varepsilon)$, then with probability at least $1 - \varepsilon$, the node u_1 will be the unique *persistent centroid* of $\mathrm{UA}(n, S_k)$, *i.e.,* for sufficiently large n, u_1 will minimize the value of ψ in $T \sim \mathrm{UA}(n, S_k)$, and remain as such throughout the rest of the attachment process. As a consequence, we can find u_1 by selecting only one node in the unlabelled tree. To summarize, when $k \geq c \log(1/\varepsilon)$,

$$\mathbf{P}\{u_1 \in H^*_{\psi;1}(T^\circ)\} \geq 1 - \varepsilon.$$

Recall also $H'_{\psi;m}$ from Section 4.

Proposition 5.2. *Let m be such that*

$$\mathbf{P}_{T\sim\mathrm{UA}(n,S_k)}\{u_1 \in H^*_{\psi;m}(T^\circ)\} \geq 1 - \varepsilon/2$$

and m' be such that

$$\mathbf{P}_{T\sim\mathrm{UA}(n,S_k)}\{L(S_k) \subseteq H'_{\psi;m'}(T^{(u_1)})\} \geq 1 - \varepsilon/2.$$

Then $K(S_k, \varepsilon) \leq mm'$.

Proof. Write $m' = K'(k, k-1, \varepsilon)$. Define

$$H = \{v \colon v \in H'_{\psi;m'}(T^{(u)}) \text{ for all } u \in H^*_{\psi;m}(T^\circ)\}.$$

Then,

$$\mathbf{P}\{V(S_k) \not\subseteq H\} \leq \mathbf{P}\{u_1 \notin H^*_{\psi;m}(T^\circ)\} + \mathbf{P}\{L(S_k) \not\subseteq H'_{\psi;m'}(T^{(u_1)})\} \leq \varepsilon$$

and clearly $|H| \leq mm'$. □

Theorem 1.7 follows from Lemma 5.1, Proposition 4.1, and Proposition 5.2.

6. Open problems

Our work raises several open problems.

1. *Joint dependence on k and ε.* From Theorem 1.1, we learn that $K^*(k, \ell, \varepsilon)$ grows roughly like $e^{1/k}$ for fixed ε, and like $\mathrm{poly}(1/\varepsilon)$ for fixed k; Theorem 1.2 tells us that $K^*(k, \ell, \varepsilon)$ grows like $e^{1/\log\log k}$ for fixed ε, and $\exp\left\{\frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)}\right\}$

for fixed k. Can we find an upper bound on $K^*(k, \ell, \varepsilon)$ which jointly behaves well as a function of ε and k, like

$$K^*(k, \ell, \varepsilon) \overset{?}{\leq} c_1 \exp\left\{ c_2 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon) + k} \right\}.$$

Should there be some dependence on ℓ?

2. *Tight bounds for K and K^*.* What is the true dependence of $K(k, \ell, \varepsilon)$ and $K^*(k, \ell, \varepsilon)$ on k, ℓ, and ε? In particular, we ask

$$K^*(k, \ell, \varepsilon) \overset{?}{\leq} c_1 \exp\left\{ c_2 \sqrt{\frac{\log(1/\varepsilon)}{k}} \right\}.$$

This question remains open even for $k = 1$, where the best and only known result is from [7]:

$$c_1 \exp\left\{ c_2 \sqrt{\log(1/\varepsilon)} \right\} \leq K(\varepsilon) \leq c_3 \exp\left\{ c_4 \frac{\log(1/\varepsilon)}{\log\log(1/\varepsilon)} \right\}.$$

3. *Lower bounds for constant ε.* Restating Conjecture 3.4, we ask: Can it be shown that for constants c, ε_0 and sufficiently large k and ℓ,

$$K(k, \ell, \varepsilon_0) \overset{?}{\geq} ck\ell.$$

4. *Partial vertex-confidence sets.* What about the optimal quantities $K^i(k, \ell, \varepsilon)$ for the smallest sets which intersect at least i nodes of seed with probability at least $1 - \varepsilon$, where $1 \leq i \leq k$? It is clear that

$$K^*(k, \ell, \varepsilon) \leq K^i(k, \ell, \varepsilon) \leq K(k, \ell, \varepsilon).$$

Can this obvious result be refined?

5. *The preferential attachment model.* Can one prove analogous upper and lower bounds on $K^i(k, \ell, \varepsilon)$ in the seeded (superlinear/sublinear) preferential attachment tree $\mathrm{UA}_\alpha(n, S)$ for $\alpha > 0$? Jog and Loh [17] showed that, for a given ε, there are c, N depending on ε such that for $K \geq N$ satisfying

$$\frac{cK(\log K)^{\frac{2}{1-\alpha}}}{(K-1)2} \leq \frac{\varepsilon}{4},$$

there exists a vertex-confidence set of size K which includes the root in $\mathrm{UA}_\alpha(n)$ with probability at least $1 - \varepsilon$.

6. *Worst-case gnostic seed recovery.* We showed how when the seed is known to be a star S_k, only a constant factor of k nodes were required to recover the seed with probability at least $1/2$. Is there any seed S for which $K(S, 1/2) = \omega_k(k)$, and in general what is the dependence on k of

$$\max_{S\colon |S|=k} K(S, 1/2)\,?$$

Any seed with $K(S, 1/2) = \omega_k(k)$ must have $\ell = \omega_k(1)$, since by (3) and Theorem 1.5,

$$K(S, 1/2) \leq K(k, \ell, 1/2) \leq ck\ell.$$

As natural candidates, we suggest that S is a complete binary tree, or a *comb graph*, namely that $V(S) = \{u_1, v_1, u_2, v_2, \ldots, u_{k/2}, v_{k/2}\}$, and

$$E(S) = \Big\{ \{u_i, u_{i+1}\}\colon 1 \leq i \leq k/2 - 1 \Big\} \cup \Big\{ \{u_i, v_i\}\colon 1 \leq i \leq k \Big\}.$$

References

[1] A. Auffinger, M. Damron, and J. Hanson. *50 Years of First-Passage Percolation*, volume 68 of *University Lecture Series*. American Mathematical Society, Providence, RI, 2017.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] S. Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Savantes Ukraine, Sect. Math.*, 1:38–49, 1924. (Russian).

[4] D. Blackwell and J. B. MacQueen. Ferguson distributions via Pólya urn schemes. *Ann. Statist.*, 1:353–355, 1973.

[5] A. Bonato. *A Course on the Web Graph*, volume 89 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI; Atlantic Association for Research in the Mathematical Sciences (AARMS), Halifax, NS, 2008.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[7] S. Bubeck, L. Devroye, and G. Lugosi. Finding Adam in random growing trees. *Random Structures Algorithms*, 50(2):158–172, 2017.

[8] S. Bubeck, R. Eldan, E. Mossel, and M. Z. Rácz. From trees to seeds: on the inference of the seed from large trees in the uniform attachment model. *Bernoulli*, 23(4A):2887–2916, 2017.

[9] S. Bubeck, E. Mossel, and M. Z. Rácz. On the influence of the seed graph in the preferential attachment model. *IEEE Trans. Network Sci. Eng.*, 2(1):30–39, 2015.

[10] N. Curien, T. Duquesne, I. Kortchemski, and I. Manolescu. Scaling limits and influence of the seed graph in preferential attachment trees. *J. Éc. polytech. Math.*, 2:1–34, 2015.

[11] L. Devroye. *Nonuniform Random Variate Generation*. Springer-Verlag, New York, 1986.

[12] L. Devroye. Applications of the theory of records in the study of random trees. *Acta Inform.*, 26(1-2):123–130, 1988.

[13] *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.0.20 of 2018-09-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller and B. V. Saunders, eds.

[14] M. Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. SpringerWienNewYork, Vienna, 2009.

[15] R. Durrett. *Random Graph Dynamics*, volume 20 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2010.

[16] J. Haigh. The recovery of the root of a tree. *J. Appl. Probability*, 7:79–88, 1970.

[17] V. Jog and P.-L. Loh. Analysis of centrality in sublinear preferential attachment trees via the Crump-Mode-Jagers branching process. *IEEE Trans. Network Sci. Eng.*, 4(1):1–12, 2017.

[18] V. Jog and P.-L. Loh. Persistence of centrality in random growing trees. *Random Structures Algorithms*, 52(1):136–157, 2018.

[19] C. Jordan. Sur les assemblages de lignes. *J. Reine Angew. Math.*, 70:185–190, 1869.

[20] J. Khim and P.-L. Loh. Confidence sets for the source of a diffusion in regular trees. *IEEE Trans. Network Sci. Eng.*, 4(1):27–40, 2017.

[21] G. Lugosi and A. S. Pereira. Finding the seed of uniform attachment trees. *ArXiv pre-print*, 2018. https://arxiv.org/abs/1801.01816.

[22] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005.

[23] J. W. Moon. The distance between nodes in recursive trees. In *Combinatorics (Proc. British Combinatorial Conf., Univ. Coll. Wales, Aberystwyth, 1973)*, pages 125–132. London Math. Soc. Lecture Note Ser., No. 13. Cambridge Univ. Press, London, 1974.

[24] H. S. Na and A. Rapoport. Distribution of nodes of a tree by degree. *Math. Biosci.*, 6:313–329, 1970.

[25] D. Shah and T. Zaman. Rumors in a network: Who's the culprit? *IEEE Trans. Inform. Theory*, 57(8):5163–5181, 2011.

[26] D. Shah and T. Zaman. Finding rumor sources on random trees. *Oper. Res.*, 64(3):736–755, 2016.

[27] P. V. Sukhatme. Tests of significance for samples of the χ^2-population with two degrees of freedom. *Ann. Hum. Genet.*, 8(1):52–56, 1937.

[28] R. van der Hofstad. *Random Graphs and Complex Networks. Vol. 1*, volume 43 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2017.

Appendix A. Supporting lemmas

Lemma A.1 (Sukhatme [11, 27])**.** *Let $U_1, \dots, U_{k-1}$ be independent identically distributed Uniform$[0,1]$ random variables, where $U_{(i)}$ denotes the i-th smallest among $U_1, \dots, U_{k-1}$. Define the spacings $S_i = U_{(i)} - U_{(i-1)}$, where $U_{(0)} = 0$ and $U_{(k)} = 1$. Then, $(S_1, \dots, S_k) \sim \mathrm{Dirichlet}(1, \dots, 1)$. Moreover, for independent identically distributed standard exponential random variables $E_1, \dots, E_k$,*

$$S_i \sim \frac{E_i}{\sum_{i=1}^k E_i} \sim \mathrm{Beta}(1, k-1)$$

for each $1 \le i \le k$, and in particular, if I is some index set of size j for $1 \le j \le k$, then

$$\min_{i \in I} S_i \sim \frac{\mathrm{Beta}(1, k-1)}{j}.$$

Lemma A.2. *Let $j \ge k$, and let T_u be the subtree of $T \sim \mathrm{UA}(n, S)$ containing the vertex labelled u after removing all edges between vertices $\{u_1, \dots, u_j\}$. Then, as $n \to \infty$,*

$$\frac{1}{n}(|T_{u_i}| \colon 1 \le i \le j) \xrightarrow{d} \mathrm{Dirichlet}(\underbrace{1, \dots, 1}_{j \text{ times}}).$$

Proof. It suffices to show that the vector $(|T_{u_i}| \colon 1 \le i \le j)$ evolves as a Pólya urn with j colours, starting with one ball of each colour, and with replacement matrix I_j, the $j \times j$ identity matrix [4]. Indeed, at each step in the attachment process wherein the node u_n is attached, it joins a subtree T_{u_i} with probability proportional to $|T_{u_i}|$ for $1 \le i \le j$, and T_{u_i} gains exactly one vertex. □

Recall that a real-valued random variable X is said to *stochastically dominate* a real-valued random variable Y if, for all $x \in \mathbb{R}$,

$$\mathbf{P}\{X \le x\} \le \mathbf{P}\{Y \le x\}.$$

Let $F_{\alpha,\beta}$ be the cumulative distribution function of a Beta(α, β) random variable, and let $f_{\alpha,\beta} : [0,1] \to \mathbb{R}$ be its density. Recall that

$$f_{\alpha,\beta}(x) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

Lemma A.3. *Let $0 < \alpha \le \beta \le \gamma$. Then,* Beta$(\beta, \gamma - \beta)$ *stochastically dominates* Beta$(\alpha, \gamma - \alpha)$.

Proof. Let $(Y_1, Y_2, Y_3) \sim \mathrm{Dirichlet}(\alpha, \beta - \alpha, \gamma - \beta)$. Then,

$$\begin{aligned} Y_1 &\sim \mathrm{Beta}(\alpha, \gamma - \alpha), \\ Y_1 + Y_2 &\sim \mathrm{Beta}(\beta, \gamma - \beta), \end{aligned}$$

so, since the former is a partial sum of the latter,

$$\mathbf{P}\{\mathrm{Beta}(\beta, \gamma - \beta) \le x\} = \mathbf{P}\{Y_1 + Y_2 \le x\} \le \mathbf{P}\{Y_1 \le x\} = \mathbf{P}\{\mathrm{Beta}(\alpha, \gamma - \alpha) \le x\}.$$

□

Lemma A.4. *Let $0 < \alpha \le \beta$. Then,* Beta$(1, \alpha)$ *stochastically dominates* Beta$(1, \beta)$.

Proof. We see, directly, for $x \in [0,1]$,

$$F_{1,\alpha}(x) = \alpha \int_0^x (1-z)^{\alpha-1}\,\mathrm{d}z = 1-(1-x)^\alpha,$$

so clearly $F_{1,\alpha}(x) \le F_{1,\beta}(x) \iff (1-x)^\alpha \ge (1-x)^\beta \iff \alpha \le \beta$. □

Lemma A.5. *Let* $0 < \ell < k$. *Then,*

$$\mathbf{P}\left\{ \left| \mathrm{Beta}(k-\ell, \ell) - \frac{k-\ell}{k} \right| \le \frac{1}{\sqrt{k}} \right\} \ge \frac{3}{4}.$$

Proof. By Chebyshev's inequality,

$$\mathbf{P}\left\{ \left| \mathrm{Beta}(k-\ell, \ell) - \frac{k-\ell}{k} \right| \ge 2\sqrt{\frac{(k-\ell)\ell}{k^2(k+1)}} \right\} \le \frac{1}{4}.$$

By the arithmetic-geometric mean inequality,

$$2\sqrt{\frac{(k-\ell)\ell}{k^2(k+1)}} \le \frac{1}{\sqrt{k+1}},$$

and the result follows. □

Email address: `lucdevroye@gmail.com, tommy.reddad@gmail.com`

School of Computer Science, McGill University, 3480 University Street, Montréal, Québec, Canada, H3A 2K6