The Annals of Applied Probability
2022, Vol. 32, No. 2, 1058–1111
https://doi.org/10.1214/21-AAP1703

CORRELATED RANDOMLY GROWING GRAPHS

BY MIKLÓS Z. RÁCZ[a] AND ANIRUDH SRIDHAR[b]

Princeton University, [a]*mracz@princeton.edu,* [b]*anirudhs@princeton.edu*

We introduce a new model of correlated randomly growing graphs and study the fundamental questions of detecting correlation and estimating aspects of the correlated structure. The model is simple and starts with any model of randomly growing graphs, such as uniform attachment (UA) or preferential attachment (PA). Given such a model, a pair of graphs (G_1, G_2) is grown in two stages: until time $t_\star$ they are grown together (i.e., $G_1 = G_2$), after which they grow independently according to the underlying growth model.

We show that whenever the seed graph has an influence in the underlying graph growth model—this has been shown for PA and UA trees and is conjectured to hold broadly—then correlation can be detected in this model, even if the graphs are grown together for just a *single time step*. We also give a general sufficient condition (which holds for PA and UA trees) under which detection is possible with probability going to 1 as $t_\star \to \infty$. Finally, we show for PA and UA trees that the amount of correlation, measured by $t_\star$, can be estimated with vanishing relative error as $t_\star \to \infty$.

CONTENTS

Received July 2020; revised May 2021.
MSC2020 subject classifications. 60C05, 05C80.
Key words and phrases. Random graphs, correlated random graphs, randomly growing graphs, influence of the seed, preferential attachment, uniform attachment.

1. Introduction. Understanding computational and inference tasks on networks is of paramount importance to solving problems in a variety of fields, including biology, sociology, and machine learning. While many of these tasks are NP-hard in the worst case, most graphs occurring in practice are not worst case, motivating the study of these problems under probabilistic generative models. Increasingly, these problems involve not just a single network but multiple networks that are correlated, and often the crux of the problem lies in understanding how the networks are correlated. Here we introduce a new model of *correlated randomly growing graphs* and study the fundamental questions of detecting correlation and estimating aspects of the correlated structure.

The model is simple and starts with any model of randomly growing graphs. A model of randomly growing graphs is specified by a *seed graph* S and a (probabilistic) *growth rule* $\mathcal{G}$ (also referred to as an *attachment rule*). We say that $\{G_t\}_{t\geq|S|}$ is a sequence of randomly growing graphs with seed S (with $|S|$ vertices) and growth rule $\mathcal{G}$, if the following two things hold. First, $G_{|S|} = S$. Subsequently, the sequence of graphs is defined inductively using $\mathcal{G}$: given G_t, the graph G_{t+1} is formed from G_t by adding a single vertex that is attached to some of the vertices in G_t, chosen according to the attachment rule $\mathcal{G}$. We write $G_n \sim \mathcal{G}(n, S)$ for an n-vertex graph generated in this way; see Figure 1 for an illustration.

For instance, an attachment rule might involve a positive integer m and the new vertex attaching to m existing vertices chosen i.i.d. according to some distribution on the existing vertices. Canonical examples include uniform attachment (UA) [19], where each existing vertex is chosen with equal probability, and preferential attachment (PA) [2, 7, 34], where each existing vertex is chosen with probability proportional to its degree. The case $m = 1$ corresponds to randomly growing trees. We write $\mathrm{UA}(n, S)$ for a UA tree on n vertices started from the seed tree S, and similarly $\mathrm{PA}(n, S)$ for a PA tree on n vertices started from S.

We are now ready to introduce the new model of *correlated* randomly growing graphs. To keep things simple, we focus on the setting of two correlated graphs. In addition to a seed graph S and a growth rule $\mathcal{G}$, the model takes an additional parameter $t_\star$, which is a positive integer satisfying $t_\star \geq |S|$. The model is simple: the two graphs G_t^1 and G_t^2 grow together until time $t_\star$, after which they grow independently. More precisely, the distribution of the sequence of the pair of graphs $\{(G_t^1, G_t^2)\}_{t\geq|S|}$ is defined as follows.

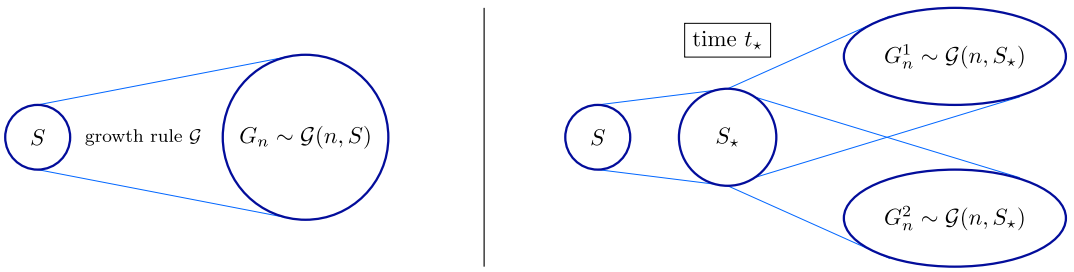
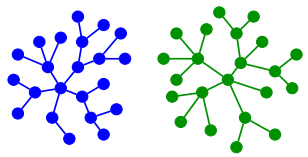
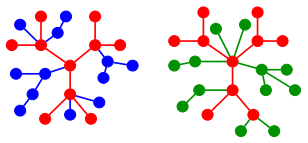


FIG. 1. *Schematic illustrations of the models studied in this paper.* Left: *a randomly growing graph, started from seed S and growing according to growth rule $\mathcal{G}$.* Right: *two correlated randomly growing graphs, started from seed S, grown together until time $t_\star$, and then growing independently.*



(a) Independent trees: $(T_n^1, T_n^2) \sim \mathcal{G}(n, S)^{\otimes 2}$.



(b) Correlated trees: $(T_n^1, T_n^2) \sim \mathcal{CG}(n, t_\star, S)$ for $t_\star = 10$, shared structure in red.

FIG. 2. *Differences between independent trees and correlated trees. Here* $n = 22$ *and* S *is the unique tree on* 3 *vertices.*

- Initially, the two graphs grow together: for $|S| \le t \le t_\star$ we have that $G_t^1 = G_t^2 =: G_t$ and $G_t \sim \mathcal{G}(t, S)$.
- Subsequently, the two graphs grow independently: conditioned on $G_{t_\star}$, the two sequences of graphs $\{G_t^1\}_{t \ge t_\star}$ and $\{G_t^2\}_{t \ge t_\star}$ are independent randomly growing graphs, both starting from the graph $G_{t_\star}$ and growing according to $\mathcal{G}$.

This can model, for instance, the citation networks [44] of two scientific fields which initially shared common beginnings but then grew apart. We write $(G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S)$ for two n-vertex graphs G_n^1 and G_n^2 generated according to this model; see Figures 1 and 2 for illustrations. We also write $\mathrm{CPA}(n, t_\star, S)$ and $\mathrm{CUA}(n, t_\star, S)$ for correlated PA trees and correlated UA trees, respectively. To the best of our knowledge, this model of correlated randomly growing graphs has not been studied before; see Section 1.3 for discussion of related work.

This model of correlation satisfies the natural property that the marginal processes are still randomly growing graphs with seed S and rule $\mathcal{G}$. That is, if $(G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S)$, then $G_n^1 \sim \mathcal{G}(n, S)$ and $G_n^2 \sim \mathcal{G}(n, S)$. Also, if $t_\star = |S|$, then $\{G_t^1\}_{t \ge |S|}$ and $\{G_t^2\}_{t \ge |S|}$ are independent; we then write $(G_n^1, G_n^2) \sim \mathcal{G}(n, S)^{\otimes 2}$ to emphasize the independence. Thus we see that $t_\star$ (more precisely, $t_\star - |S|$) explicitly measures the amount of correlation among the two graphs.

1.1. *Questions*: *Detection and estimation.* We study the fundamental questions of detecting correlation and estimating aspects of the correlated structure in the model of correlated randomly growing graphs introduced above.

Detection. Given two (unlabeled) n-vertex graphs, G_n^1 and G_n^2, can we detect whether they are correlated or not? This question can be phrased as a simple hypothesis testing problem. Under the null hypothesis H_0, the two graphs are independent: $(G_n^1, G_n^2) \sim \mathcal{G}(n, S)^{\otimes 2}$. Under the alternative hypothesis, denoted $H_{t_\star}$, the two graphs are correlated, with a shared history until time $t_\star$: $(G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S)$. In brief,

$$H_0 : (G_n^1, G_n^2) \sim \mathcal{G}(n, S)^{\otimes 2}, \qquad H_{t_\star} : (G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S). \tag{1.1}$$

Note that we only observe a *snapshot* of the two graphs at time n, we do not observe their history leading up to this snapshot. Is there a test that can distinguish between the two hypotheses with asymptotically (in n) nonnegligible power? Under what circumstances can we distinguish with probability close to 1? Studying these questions is equivalent to understanding the total variation distance between $\mathcal{G}(n, S)^{\otimes 2}$ and $\mathcal{CG}(n, t_\star, S)$; recall that the total variation distance between two probability measures P and Q is defined as $\mathrm{TV}(P, Q) := \frac{1}{2}\|P - Q\|_1 = \sup_A |P(A) - Q(A)|$. We are particularly interested in the limit as $n \to \infty$:

$$\lim_{n \to \infty} \mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big), \tag{1.2}$$

a limit which is well defined, because this total variation distance is nonincreasing in n (since one can simulate the future evolution of the process) and nonnegative. There exists a test with asymptotically nonnegligible power for the hypothesis testing problem in (1.1) if and only if the quantity in (1.2) is positive.

Estimation. If detection is possible, the natural next questions concern estimation. Is it possible to estimate the amount of correlation between two correlated randomly growing graphs? Is it possible to estimate the common shared subgraph? Formally, suppose that $(G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S)$, but $t_\star$ is unknown. How well can we estimate $t_\star$? How well can we estimate the shared subgraph $G_{t_\star}$?

1.2. *Summary of results and methods*. Our results concern the detection and estimation questions discussed in Section 1.1, and can be summarized as follows.

- *Detecting correlation whenever the seed has an influence.* We show that there exists a test with asymptotically (in n) nonnegligible power for the hypothesis testing problem in (1.1) whenever the seed graph S has an influence on the randomly growing graph $\mathcal{G}(n, S)$ (in a sense to be made precise). This latter property has been shown for PA trees [10, 16] and UA trees [9]—and is conjectured to hold more broadly—which implies that detecting correlation is possible for these models. Remarkably, the results show that correlation can be detected whenever $t_\star > |S|$, that is, even if the graphs are grown together for just a *single time step*.
- *Detecting correlation with probability going to* 1 *as* $t_\star \to \infty$. We give a general condition under which correlation can be detected with probability going to 1 as $t_\star \to \infty$. We conjecture that this condition holds for a broad family of randomly growing graphs, and in particular, we show that it holds for PA and UA trees.
- *Estimating* $t_\star$ *with vanishing relative error as* $t_\star \to \infty$. Focusing on PA and UA trees, we show that the amount of correlation, measured by $t_\star$, can be estimated with vanishing relative error as $t_\star \to \infty$.

In the most general setting, we establish results for sequential attachment rules that are *Markov*, in the sense that for every $t \geq |S|$, we have that

$$\mathbb{P}(G_{t+1} = G \mid G_{|S|}, G_{|S|+1}, \ldots, G_t) = \mathbb{P}(G_{t+1} = G \mid G_t),$$

where $\{G_t\}_{t \geq |S|}$ is a sequence of randomly growing graphs starting from seed S. This is a natural assumption, since in many real-world networks new nodes added to the network will not have access to the history of the network. We also establish stronger results for PA and UA trees, which are canonical models of randomly growing graphs. For what follows it will be useful to define

$$\text{Range}(\mathcal{G}, S) := \big\{G : \exists n \text{ such that if } G_n \sim \mathcal{G}(n, S) \text{ then } \mathbb{P}(G_n = G) > 0\big\},$$

the set of all possible graphs that can be obtained with positive probability starting from seed graph S via the attachment rule $\mathcal{G}$. We are now ready to detail our results.

1.2.1. *Detecting correlation whenever the seed has an influence*. Our first result is a general result that shows that correlation can be detected whenever the seed graph has an influence in the underlying randomly growing graph model.

THEOREM 1.1 (Detecting correlation whenever the seed has an influence). *Fix a seed graph* S, *a positive integer* $t_\star$ *such that* $t_\star > |S|$, *and a Markov sequential attachment rule* $\mathcal{G}$. *Suppose that there are graphs* G *and* G' *satisfying that* $|G| = |G'| = t_\star$, *that* $G, G' \in \text{Range}(\mathcal{G}, S)$, *and that*

$$\lim_{n \to \infty} \text{TV}\big(\mathcal{G}(n, G), \mathcal{G}(n, G')\big) > 0. \tag{1.3}$$

Then

$$\lim_{n \to \infty} \text{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) > 0.$$

Remarkably, this result holds whenever $t_\star > |S|$, showing that correlation can be detected even if the graphs are grown together for just a single time step.

The condition in (1.3) captures formally what it means for the seed to have an influence. The study of the influence of the seed in randomly growing graphs was initiated by Bubeck, Mossel, and Rácz, who studied this question in PA trees [10]. They showed that for any two seed trees S and T with at least 3 vertices and different degree profiles, $\lim_{n\to\infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T)) > 0$ holds. This already implies that (1.3) holds for PA trees whenever $t_\star > 3$. In subsequent work, Curien, Duquesne, Kortchemski, and Manolescu showed that $\lim_{n\to\infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T)) > 0$ whenever S and T are nonisomorphic trees with at least 3 vertices [16]. This was then showed for UA trees as well by Bubeck, Eldan, Mossel, and Rácz [9]. We refer to the recent survey [42] for an exposition of these results and the associated techniques. These results are summarized in the following two theorems.

THEOREM 1.2 ([10, 16]). *The seed has an influence in PA trees in the following sense. We have that* $\lim_{n\to\infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T)) > 0$ *for any trees S and T that are nonisomorphic and have at least* 3 *vertices.*

THEOREM 1.3 ([9]). *The seed has an influence in UA trees in the following sense. We have that* $\lim_{n\to\infty} \mathrm{TV}(\mathrm{UA}(n, S), \mathrm{UA}(n, T)) > 0$ *for any trees S and T that are nonisomorphic and have at least* 3 *vertices.*

These two theorems, together with Theorem 1.1, directly imply that correlation can be detected in PA and UA trees. These results are formalized in the following two corollaries.

COROLLARY 1.4 (Detecting correlation in PA trees). *Let S be a finite tree with at least two vertices. Let $t_\star \in \mathbb{N}$ be such that $t_\star > |S|$ and $t_\star > 3$. Then*

$$\lim_{n\to\infty} \mathrm{TV}\big(\mathrm{CPA}(n, t_\star, S), \mathrm{PA}(n, S)^{\otimes 2}\big) > 0.$$

COROLLARY 1.5 (Detecting correlation in UA trees). *Let S be a finite tree. Let $t_\star \in \mathbb{N}$ be such that $t_\star > |S|$ and $t_\star > 3$. Then*

$$\lim_{n\to\infty} \mathrm{TV}\big(\mathrm{CUA}(n, t_\star, S), \mathrm{UA}(n, S)^{\otimes 2}\big) > 0.$$

Theorem 1.1 reduces detecting correlation to detecting the influence of the seed. As such, it can be viewed as an *existence* result, since it does not give specific statistics of the two graphs that can detect correlation. We therefore complement Theorem 1.1 and Corollaries 1.4 and 1.5 by providing alternative, *algorithmic* proofs of Corollary 1.4 and Corollary 1.5. Specifically, inspired by [10], we will show that the maximum degrees of the two trees can be used to detect correlation in PA trees. Furthermore, inspired by [9], we will show that there are certain statistics that measure global balancedness properties of a tree (and which are efficiently computable) that can be used to detect correlation in UA trees. See the Supplementary Material [43] for details.

1.2.2. *Detecting correlation with probability going to* 1 *as* $t_\star \to \infty$. Ideally, we would like to detect correlation with probability close to 1. However, for any fixed finite $t_\star$, the probability of successfully being able to detect correlation is strictly bounded away from 1. This is simply because if $G^1_{t_\star} \sim \mathcal{G}(t_\star, S)$ and $G^2_{t_\star} \sim \mathcal{G}(t_\star, S)$ are independent, then there is a positive probability (which depends only on $\mathcal{G}$ and $t_\star$) that $G^1_{t_\star} = G^2_{t_\star}$. With this probability we may couple $\mathcal{G}(n, S)^{\otimes 2}$ and $\mathcal{CG}(n, t_\star, S)$, showing that there exists $\varepsilon = \varepsilon(\mathcal{G}, t_\star) > 0$ such

that

$$\mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) \le 1 - \varepsilon \tag{1.4}$$

for every $n \ge t_\star$. Our focus is thus to show that correlation can be detected with probability going to 1 as $t_\star \to \infty$. We first present a general result, which gives a sufficient condition on the underlying model of randomly growing graphs for this to occur.

THEOREM 1.6 (Detecting correlation with probability going to 1 as $t_\star \to \infty$). *Fix a seed graph S and a Markov sequential attachment rule $\mathcal{G}$. Let $\{G_t\}_{t \ge |S|}$ be a sequence of randomly growing graphs with seed S and attachment rule $\mathcal{G}$. Suppose that there is a function $f : \mathrm{Range}(\mathcal{G}, S) \to \mathbb{R}$ such that the limit $\lim_{t\to\infty} f(G_t) =: f_\infty$ exists almost surely and that f_∞ is an absolutely continuous random variable with respect to the Lebesgue measure. Then we have that*

$$\lim_{t_\star\to\infty} \lim_{n\to\infty} \mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) = 1.$$

The test that distinguishes correlated graphs from independent graphs is simple: we compare $|f(G_n^1) - f(G_n^2)|$ to an appropriately chosen threshold. The idea behind the proof is that this quantity tends to 0 as $t_\star \to \infty$ under the alternative hypothesis $H_{t_\star}$, but $f(G_n^1)$ and $f(G_n^2)$ are independent under the null hypothesis H_0, so the difference stays away from 0 in this case.

Theorem 1.6 is a general theorem that we expect applies to a wide class of models of randomly growing graphs. To demonstrate its utility, we show that PA trees and UA trees satisfy its conditions. For PA trees, we may choose f to be the normalized maximum degree. For both cases, we may choose f to be a function that is closely related to notions of centrality in trees. These have been used to study a variety of statistical problems, such as estimating the source of a rumor on a tree [45–47] and estimating the seed in randomly growing trees [8, 17, 32]. We thus obtain the following results for PA and UA trees.

THEOREM 1.7. *Let S be a finite tree with at least two vertices. Then*

$$\lim_{t_\star\to\infty} \lim_{n\to\infty} \mathrm{TV}\big(\mathrm{CPA}(n, t_\star, S), \mathrm{PA}(n, S)^{\otimes 2}\big) = 1.$$

THEOREM 1.8. *Let S be a finite tree. Then*

$$\lim_{t_\star\to\infty} \lim_{n\to\infty} \mathrm{TV}\big(\mathrm{CUA}(n, t_\star, S), \mathrm{UA}(n, S)^{\otimes 2}\big) = 1.$$

1.2.3. *Estimating $t_\star$ with vanishing relative error as $t_\star \to \infty$.* We now turn to questions of estimation. These are more involved than questions concerning detection and hence we restrict our attention to PA and UA trees, started from the seed $S = S_2$, the unique tree on two vertices. We focus on estimating $t_\star$, which measures the amount of correlation between the two correlated trees; we leave the very interesting question of estimating the common subgraph $G_{t_\star}$ for future work (see Section 1.4). Ideally, we would like good estimates of $t_\star$ that hold with probability close to 1. From (1.4) it follows that this is only possible as $t_\star \to \infty$.

Our main result on estimation is that $t_\star$ can be estimated with vanishing relative error as $t_\star \to \infty$; this is the content of the following theorem.

THEOREM 1.9 (Estimating $t_\star$ in PA and UA trees). *Let $S = S_2$ be the unique tree on two vertices and let $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. There exists an estimator $\widehat{t}_n \equiv \widehat{t}(T_n^1, T_n^2)$, computable in polynomial time, such that*

$$\lim_{t_\star\to\infty} \liminf_{n\to\infty} \mathbb{P}\left(\left(1 - \frac{\log\log t_\star}{\sqrt{\log t_\star}}\right)t_\star \le \widehat{t}_n \le \left(1 + \frac{\log\log t_\star}{\sqrt{\log t_\star}}\right)t_\star\right) = 1.$$

The same result also holds when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

In other words, the relative error of the estimator $\widehat{t}_n$ is bounded by $\log\log(t_\star)/\sqrt{\log t_\star}$, with probability close to 1, for large enough $t_\star$. The proof of Theorem 1.9 is the most involved proof in this paper and so we give here a high level overview of the proof strategy. The proof works equally for both PA and UA trees, with only minor changes.

The main idea is to match several pairs of vertices across the two trees. To explain this more precisely, we introduce some notation. Let $\{T_n\}_{n\geq 2}$ be a sequence of growing trees with seed S_2. For a vertex v in T_n, let $\tau(v)$ be the *timestamp* of v. That is, $\tau(v) = k$ if v is not in T_{k-1} but is introduced in T_k. The two initial vertices are labeled 1 and 2 arbitrarily. We say that a pair of vertices (v^1, v^2), where $v^1 \in V(T_n^1)$ and $v^2 \in V(T_n^2)$, is *correctly matched* if $\tau(v^1) = \tau(v^2)$.

Correctly matching the centroids. Let $\theta^1(n), \theta^2(n)$ be the centroids of the trees T_n^1 and T_n^2, respectively (we rigorously define the notion of a tree centroid in Section 3.2). Jog and Loh [27] proved that PA and UA trees with seed S_2 have the *persistent centroid property*: almost surely, there is a finite time N such that for all $t \geq N$, we have that $\theta(t) = \theta(N) := \theta$. Using this fact, it follows that the pair $(\theta^1(n), \theta^2(n))$ is correctly matched with probability tending to 1 as $t_\star \to \infty$. Although we have so far only matched one pair of vertices in the two graphs, this provides an important frame of reference going forward, to analyze the correlated structure in the two trees.

Matching neighbors of the centroids. Next, assuming the high-probability event $\theta^1(n) = \theta^2(n) = \theta$, we consider the rooted trees (T_n^1, θ) and (T_n^2, θ), with the goal of matching many neighbors of the centroids. We do so by examining subtrees of the two rooted trees. Let $(T_n^i, \theta)_{v\downarrow}$ denote the subtree of the rooted tree (T_n^i, θ) that has root v. In other words, the tree $(T_n^i, \theta)_{v\downarrow}$ consists of all vertices u such that the unique path connecting u and θ passes through v.

The idea behind matching neighbors of the centroid is the "rich-get-richer" property of subtrees. To illustrate this concept, suppose that for a tree growing via uniform attachment, we consider neighbors u and v of θ, and $|(T_{t_\star}, \theta)_{u\downarrow}|$ is much larger than $|(T_{t_\star}, \theta)_{v\downarrow}|$. Under the UA rule, the probability that a new vertex joins a subtree is proportional to the number of vertices in the subtree; thus it is very unlikely that $|(T_t, \theta)_{v\downarrow}|$ exceeds $|(T_t, \theta)_{u\downarrow}|$ at any future time t. Similar behavior holds for PA trees as well. This intuition tells us that if $|(T_{t_\star}, \theta)_{u\downarrow}|$ is much larger than $|(T_{t_\star}, \theta)_{v\downarrow}|$, then we should have $|(T_n^i, \theta)_{u\downarrow}| > |(T_n^i, \theta)_{v\downarrow}|$ for both $i = 1$ and $i = 2$.

Taking this idea one step further, we may expect that if the largest R subtrees (for some positive integer R) of $(T_{t_\star}, \theta)$ do not have sizes that are too close to each other, then these should be the same R largest subtrees in (T_n^i, θ), for both $i = 1$ and $i = 2$. Therefore, we will match the neighbors of the centroids with the largest subtrees, the second largest subtrees, and so on, until the Rth largest subtrees. We indeed prove that such a matching procedure for the neighbors of the centroids, based on subtree ranking, gives us all correct matchings with probability tending to 1 as $t_\star \to \infty$.

Constructing estimators for $t_\star$. Suppose that (v^1, v^2) are a correctly matched pair of neighbors of the centroid. We can construct an estimator for $t_\star$ by comparing the subtree sizes corresponding to v^1 and v^2. The evolution of subtree sizes in PA and UA trees exhibit the following stability property: the fraction of vertices that lie in a particular subtree has a limit almost surely as the size of the tree tends to infinity. This follows from viewing the subtree growth as a Pólya urn process.

We then expect that as we send $t_\star \to \infty$, the difference between $\frac{1}{n}|(T_n^1, \theta)_{v^1\downarrow}|$ and $\frac{1}{n}|(T_n^2, \theta)_{v^2\downarrow}|$ is close to 0, even for large n. We exploit this property to construct a nearly unbiased estimator for $t_\star$ based on the difference between $\frac{1}{n}|(T_n^1, \theta)_{v^1\downarrow}|$ and $\frac{1}{n}|(T_n^2, \theta)_{v^2\downarrow}|$.

However, the variance of the estimator corresponding to the matched pair (v^1, v^2) is not small enough to ensure that we can estimate $t_\star$ with vanishing relative error. This is the reason for matching many pairs of points: we can then average the estimators corresponding to many correctly matched pairs of vertices, in order to reduce the variance. We finish by applying Chebyshev's inequality.

1.3. *Related work.* Though this paper is, to the best of our knowledge, the first to introduce this model of correlated randomly growing graphs, it is closely related to several well-studied problems in the literature.

1.3.1. *Graph matching and the correlated Erdős–Rényi model.* Perhaps the most well-known related problem is graph matching. In this setting, we are given two graphs and we want to find a labeling on the vertices that maximizes the similarity between the two graphs. The applications of this problem are numerous, spanning data privacy in social networks [37, 41], protein-protein interaction networks [48], computer vision [11], pattern recognition [4, 12], machine learning [13], and more. This problem is NP-hard in the worst case (see, e.g., the surveys [12, 31]); in fact, it is even hard to approximate under some hardness assumptions [39]. However, most graphs occurring in applications are not worst case, which motivates the study of the graph matching problem under probabilistic generative models.

The simplest random graph model is the Erdős–Rényi random graph $G(n, p)$, which has n vertices and every pair is connected with probability p, independently of any other pair. Thus naturally the simplest model of correlated random graphs involves two Erdős–Rényi random graphs that are correlated. This model was introduced by Pedarsani and Grossglauser [41] and has been widely studied in the past decade in several communities, including computer science, network science, information theory, probability, and statistics [3, 14, 15, 18, 20, 21, 25, 28–30, 33, 36, 49]. These works have resulted in obtaining the fundamental information-theoretic limits [14, 15] and recent algorithmic advances [3, 18, 20, 21, 36]. The model of correlated randomly grown graphs introduced in this paper is fundamentally different from the correlated Erdős–Rényi model and thus it is not possible to directly compare our results with those in these papers. Importantly, while Erdős–Rényi random graphs have no inherent structure, the model of correlated randomly grown graphs is motivated by the fact that many real-world networks form via a growth process.

In the correlated Erdős–Rényi model the pair (G^1, G^2) is constructed as follows. First, sample an unobserved base graph $G^0 \sim G(n, p)$. Next, conditioned on G^0, construct G^1 and G^2 independently by including any given edge with probability q. Both G^1 and G^2 are distributed according to $G(n, pq)$, and they are correlated in the sense that the presence of specified edges are correlated. There is also a "true" labeling of the vertices in G^1 and G^2, given by inheriting the labels of the unobserved base graph G^0. The goal of the graph matching problem is to recover this true labeling (up to isomorphism). There is also a modified version of the problem in which the algorithm has side information in the form of a small number of matched vertices.

The problems of detecting and estimating correlation in a pair of randomly grown graphs can be viewed as an analog of the graph matching problem (without side information) for these kind of graphs. We highlight several papers in the graph matching literature that have related ideas. Barak, Chou, Lei, Schramm, and Sheng study the problem of detecting correlated structure for a pair of Erdős–Rényi graphs [3]. Their approach to solving the detection problem in certain regimes relies on subgraph counts. Our approach is vastly different, relying on extremal statistics of the graphs (e.g., maximum degree, minimum anti-centrality) and general balancedess properties (all of which may be computed efficiently). Kazemi, Yartseva, and Grossglauser study a variant of the graph matching problem in a pair of correlated

Erdős–Rényi graphs when there is *partial overlap* between the graphs; that is, there are vertices in either graph that are not part of any correlated structure [29]. Our model of correlated randomly grown graphs has a similar characteristic: the subgraph of the shared history, $G_{t_\star}$, is common, and the other vertices in the pair of graphs do not necessarily correspond to each other if they were born after time $t_\star$. Their goal is somewhat different from ours; they aim to estimate the common part, with knowledge of the amount of overlap. On the other hand, we focus on estimating the amount of correlation, or equivalently, the size of the common part.

Korula and Lattanzi study a version of the graph matching problem for preferential attachment graphs [30], though the manner in which they generate a pair of correlated graphs is fundamentally different from our model. Similar to the process of generating correlated Erdős–Rényi graphs, they generate a base graph G^0 according to preferential attachment and independently construct G^1 and G^2 by including a given edge in G^0 with some fixed probability. However, in this case G^1 and G^2 are not distributed according to preferential attachment, which is unnatural. We also note that they require the use of side information in their algorithm, while we do not assume this, since it is possible to match key information in our case (e.g., matching the centroid).

1.3.2. *Inferring the history of a dynamic graph process from a snapshot.* Our work naturally fits under this broad category in terms of the problem scope and the techniques used. There have been a variety of works of this theme in recent years, including rumor source estimation [22–24, 45–47], the influence of the seed in randomly growing graphs [9, 10, 16], and finding the earliest vertices in randomly growing graphs [8, 17, 32]. Applications include reconstructing the evolution of biological networks [38].

The works on the influence of the seed in randomly growing graphs [9, 10, 16] are particularly relevant to our work—we refer to Section 1.2 for a discussion of these detailed connections. These connections are further touched upon in the proofs.

The notion of centrality in trees plays a significant role in our techniques (for the results specific to PA and UA trees), and in many of the cited works. Shah and Zaman formulated the notion of *rumor centrality* for maximum likelihood estimation of the source of a diffusion on a tree [45–47]. Bubeck, Devroye, and Lugosi introduced a related centrality measure based on subtree sizes to obtain confidence intervals for the first vertex in a PA or UA tree [8]. This centrality measure lends itself to an easier analysis with PA and UA trees, since the evolution of subtree sizes can be understood as Pólya urn processes. Subsequently, this centrality measure was used by Lugosi and Pereira [32] and by Devroye and Reddad [17] for the more general problem of obtaining confidence intervals for the seed graph of a UA tree, as well as for the earliest vertices. Jog and Loh showed that UA trees and PA trees exhibit the persistent centroid property: the location of the centroid (with respect to the centrality measure of [8]) only changes finitely many times as the number of vertices in the tree increases [26, 27]. We are able to leverage these previous results on centrality in our study of the detection and estimation problems for PA and UA trees.

Bhamidi, Jin, and Nobel studied a variant of the preferential attachment model with a change point [6] (see also [1])—this shares some similar elements to our model but is fundamentally different. In their model, they examine a single PA tree where, at some time point, the attachment rule changes. The goal is to estimate this change point, and to do so, they use knowledge of the history of the graph. Our problem can be viewed as a change point problem as well, but in a much different sense. Both of the randomly grown graphs have the marginal distribution of a standard randomly grown graph, and the correlation time $t_\star$ may be interpreted as a change point when the two growing graphs begin to evolve independently. Also, we observe a single snapshot, rather than the entire history, which is a more appropriate and interesting setting for our problem.

Finally, there are many important aspects of modeling network formation that are beyond the scope of the present article. We refer the reader to the recent work of Overgoor, Benson, and Ugander [40], which unifies a host of network formation models using a framework based on discrete choice theory. (See also the references therein for an overview of the related literature.) Our hope is that the novel phenomena presented in this article can contribute to the broader discussion on modeling the formation of multiple correlated networks.

1.4. *Discussion and open problems.* This paper initiates the study of correlated randomly growing graphs and leaves open several problems. We end the Introduction by discussing possible future directions.

- *Estimating the correlation time $t_\star$.* We have shown (in PA and UA trees) that the correlation time $t_\star$ can be estimated with vanishing relative error as $t_\star \to \infty$. It would be interesting to understand the limits of how well $t_\star$ can be estimated.
- *Estimating the common subgraph $G_{t_\star}$.* It is of great interest to estimate the common subgraph $G_{t_\star}$ shared by the two correlated randomly growing graphs. This question can be formalized in several ways: for instance, we might want to find a large subgraph of $G_{t_\star}$ or a small supergraph of $G_{t_\star}$, with probability close to 1. Recent work by Lugosi and Pereira [32] and Devroye and Reddad [17] (following work by Bubeck, Devroye, and Lugosi [8]) has studied seed-finding algorithms for UA trees. We suspect that their results and the techniques they have developed will be useful for estimating $G_{t_\star}$.
- *Other models of randomly growing graphs.* In our work we focus on PA and UA trees when studying specific models of randomly growing graphs. Our general result in Theorem 1.1 says that correlation can be detected if (1.3) holds. This is a much weaker form of the influence of the seed than is established in Theorems 1.2 and 1.3 for PA and UA trees. Are there models of randomly growing graphs for which it is possible to show that (1.3) holds even if showing the analogue of Theorems 1.2 and 1.3 is currently out of reach?
- *Large amounts of correlation.* In our work we have focused on $t_\star$ being fixed compared to the graph size n. What if $t_\star$ is a function of n? This introduces much more correlation among the two graphs and it would be interesting to understand how much stronger results can be obtained.
- *Three or more correlated graphs.* The introduced model of correlated randomly growing graphs naturally extends to three or more correlated graphs. How do the questions of detection and estimation change in this setting? For instance, is it much easier to estimate the common subgraph $G_{t_\star}$ if we have samples from many correlated graphs?

1.5. *Outline.* The rest of the paper is organized as follows. We start with proving Theorem 1.1 in Section 2. In Section 3 we turn to detecting correlation with probability going to 1 as $t_\star \to \infty$ and prove Theorems 1.6, 1.7, and 1.8. Finally, we turn to estimating $t_\star$ as $t_\star \to \infty$. We first provide an initial, coarse estimate of $t_\star$ in Section 4; this section contains the main ideas of our estimators. However, further ideas are needed in order to obtain an estimator of $t_\star$ which has vanishing relative error as $t_\star \to \infty$: these, and a proof of Theorem 1.9, can be found in Section 5.

Proofs of certain technical lemmas in Sections 4 and 5 are deferred to the Appendix. In addition, we present explicit algorithmic proofs of Corollaries 1.4 and 1.5 in the Supplementary Material.

2. Detecting correlation when the seed has an influence. In this section we prove Theorem 1.1. To abbreviate notation, in the following we denote by $\mathbb{P}_0$ the underlying probability measure when $(G_n^1, G_n^2) \sim \mathcal{G}(n, S)^{\otimes 2}$ and by $\mathbb{P}_{t_\star}$ the underlying probability measure when

$(G_n^1, G_n^2) \sim \mathcal{CG}(n, t_\star, S)$. Furthermore, for a graph H we denote by $\mathbb{P}_H$ the probability measure on the sequence of randomly growing graphs $\{G_n\}_{n \geq |H|}$ with seed H and attachment rule $\mathcal{G}$.

PROOF OF THEOREM 1.1. From (1.3) it follows that there exist $\delta > 0$ and a sequence $\{\mathcal{E}_n\}_{n \geq t_\star}$ such that

$$\big|\mathbb{P}_G(G_n \in \mathcal{E}_n) - \mathbb{P}_{G'}(G_n \in \mathcal{E}_n)\big| \geq \delta \tag{2.1}$$

for every $n \geq t_\star$. Turning now to a pair of graphs (G_n^1, G_n^2), with $n \geq t_\star$, we consider the event

$$\{G_n^1 \in \mathcal{E}_n\} \cap \{G_n^2 \in \mathcal{E}_n\}.$$

Under the null hypothesis H_0, the two graphs G_n^1 and G_n^2 are independent, and thus we have that

$$\mathbb{P}_0\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) = \mathbb{P}_0\big(G_n^1 \in \mathcal{E}_n\big)\mathbb{P}_0\big(G_n^2 \in \mathcal{E}_n\big) = \big(\mathbb{P}_S(G_n \in \mathcal{E}_n)\big)^2.$$

Note also that by conditioning on the graph at time $t_\star$ and using the fact that the sequential attachment rule $\mathcal{G}$ is Markov, we have that

$$\mu := \mathbb{P}_S(G_n \in \mathcal{E}_n) = \sum_{H:|H|=t_\star} \mathbb{P}_H(G_n \in \mathcal{E}_n)\mathbb{P}_S(G_{t_\star} = H),$$

where the sum is over all graphs on $t_\star$ vertices.

Turning to the alternative hypothesis $H_{t_\star}$, we can again condition on the graph at time $t_\star$, and use the fact G_n^1 and G_n^2 are independent conditioned on the graph at time $t_\star$. We thus obtain that

$$\begin{aligned}
&\mathbb{P}_{t_\star}\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) \\
&\quad = \sum_{H:|H|=t_\star} \mathbb{P}_{t_\star}\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n \mid G_{t_\star}^1 = G_{t_\star}^2 = H\big)\mathbb{P}_S(G_{t_\star} = H) \\
&\quad = \sum_{H:|H|=t_\star} \mathbb{P}_{t_\star}\big(G_n^1 \in \mathcal{E}_n \mid G_{t_\star}^1 = G_{t_\star}^2 = H\big)\mathbb{P}_{t_\star}\big(G_n^2 \in \mathcal{E}_n \mid G_{t_\star}^1 = G_{t_\star}^2 = H\big)\mathbb{P}_S(G_{t_\star} = H) \\
&\quad = \sum_{H:|H|=t_\star} \big(\mathbb{P}_H(G_n \in \mathcal{E}_n)\big)^2\mathbb{P}_S(G_{t_\star} = H).
\end{aligned}$$

Altogether, we have thus obtained that

$$\begin{aligned}
&\mathbb{P}_{t_\star}\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) - \mathbb{P}_0\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) \\
&\quad = \sum_{H:|H|=t_\star} \big(\mathbb{P}_H(G_n \in \mathcal{E}_n)\big)^2\mathbb{P}_S(G_{t_\star} = H) - \Big(\sum_{H:|H|=t_\star} \mathbb{P}_H(G_n \in \mathcal{E}_n)\mathbb{P}_S(G_{t_\star} = H)\Big)^2 \\
&\quad = \sum_{H:|H|=t_\star} \mathbb{P}_S(G_{t_\star} = H)\big(\mathbb{P}_H(G_n \in \mathcal{E}_n) - \mu\big)^2.
\end{aligned}$$

Note that all terms in this sum are nonnegative. Dropping all terms except those corresponding to G and G', we have that

$$\begin{aligned}
&\mathbb{P}_{t_\star}\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) - \mathbb{P}_0\big(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n\big) \\
&\qquad \geq \mathbb{P}_S(G_{t_\star} = G)\big(\mathbb{P}_G(G_n \in \mathcal{E}_n) - \mu\big)^2 + \mathbb{P}_S\big(G_{t_\star} = G'\big)\big(\mathbb{P}_{G'}(G_n \in \mathcal{E}_n) - \mu\big)^2.
\end{aligned}$$

By the condition that $G, G' \in \mathrm{Range}(\mathcal{G}, S)$, we have that $\mathbb{P}_S(G_{t_\star} = G)$ and $\mathbb{P}_S(G_{t_\star} = G')$ are both strictly positive, and note that these are not a function of n. By (2.1) it follows that at least

one of $\mathbb{P}_G(G_n \in \mathcal{E}_n)$ and $\mathbb{P}_{G'}(G_n \in \mathcal{E}_n)$ must be outside of the interval $(\mu - \delta/2, \mu + \delta/2)$, showing that

$$\big(\mathbb{P}_G(G_n \in \mathcal{E}_n) - \mu\big)^2 + \big(\mathbb{P}_{G'}(G_n \in \mathcal{E}_n) - \mu\big)^2 \geq \delta^2/4.$$

Putting everything together, we have shown that

$$\begin{aligned}&\mathbb{P}_{t_\star}(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n) - \mathbb{P}_0(G_n^1 \in \mathcal{E}_n, G_n^2 \in \mathcal{E}_n)\\ &\quad\geq \frac{\delta^2}{4}\min\big\{\mathbb{P}_S(G_{t_\star} = G), \mathbb{P}_S\big(G_{t_\star} = G'\big)\big\}\end{aligned}$$

for every $n \geq t_\star$, which implies that

$$\lim_{n\to\infty} \mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) \geq \frac{\delta^2}{4}\min\big\{\mathbb{P}_S(G_{t_\star} = G), \mathbb{P}_S\big(G_{t_\star} = G'\big)\big\} > 0. \qquad \square$$

3. Detecting correlation with probability going to 1 as $t_\star \to \infty$. In this section we focus on detecting correlation with probability going to 1 as $t_\star \to \infty$. We first prove Theorem 1.6 in Section 3.1 and then prove Theorems 1.7 and 1.8 in Section 3.2.

3.1. *A sufficient condition for Markov sequential attachment rules.* To abbreviate notation, in the following we denote by $\mathbb{P}_0$ the underlying probability measure when $\{(G_t^1, G_t^2)\}_{t\geq|S|}$ are two independent sequences of randomly growing graphs with seed S and attachment rule $\mathcal{G}$. Similarly, we denote by $\mathbb{P}_{t_\star}$ the underlying probability measure when the two graphs are correlated until time $t_\star$. We say that a random variable is *absolutely continuous* when its distribution is absolutely continuous with respect to the Lebesgue measure.

PROOF OF THEOREM 1.6. We will show that for every $\delta > 0$ there exists $t' = t'(\delta)$ such that for every $t_\star \geq t'$ we have that

$$\lim_{n\to\infty} \mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) \geq 1 - \delta. \tag{3.1}$$

To this end, fix $\delta > 0$. Let $\{(G_t^1, G_t^2)\}_{t\geq|S|}$ be two sequences of randomly growing graphs with seed S and attachment rule $\mathcal{G}$, under either $\mathbb{P}_0$ or $\mathbb{P}_{t_\star}$. Let $f_\infty^1 := \lim_{t\to\infty} f(G_t^1)$ and $f_\infty^2 := \lim_{t\to\infty} f(G_t^2)$; by our assumptions these limits exist almost surely, under both $\mathbb{P}_0$ and $\mathbb{P}_{t_\star}$. Observe that

$$\lim_{\varepsilon\to 0} \mathbb{P}_0\big(|f_\infty^1 - f_\infty^2| \leq \varepsilon\big) = \mathbb{P}_0\big(f_\infty^1 = f_\infty^2\big) = 0,$$

the latter equality holding because f_∞^1 and f_∞^2 are i.i.d. absolutely continuous random variables under $\mathbb{P}_0$. Thus fix $\varepsilon > 0$ such that

$$\mathbb{P}_0\big(|f_\infty^1 - f_\infty^2| \leq \varepsilon\big) \leq \delta/2. \tag{3.2}$$

Turning to the measure $\mathbb{P}_{t_\star}$, note that under $\mathbb{P}_{t_\star}$ we have that $G_{t_\star}^1 = G_{t_\star}^2$ almost surely, and hence $f(G_{t_\star}^1) = f(G_{t_\star}^2)$ almost surely as well. So by the triangle inequality we have, for any $n \geq t_\star$, that

$$\begin{aligned}&\mathbb{P}_{t_\star}\big(|f(G_n^1) - f(G_n^2)| > \varepsilon\big)\\ &\quad\leq \mathbb{P}_{t_\star}\big(|f(G_n^1) - f(G_{t_\star}^1)| > \varepsilon/2\big) + \mathbb{P}_{t_\star}\big(|f(G_n^2) - f(G_{t_\star}^2)| > \varepsilon/2\big)\\ &\quad= 2\mathbb{P}_{t_\star}\big(|f(G_n^1) - f(G_{t_\star}^1)| > \varepsilon/2\big) = 2\mathbb{P}_0\big(|f(G_n^1) - f(G_{t_\star}^1)| > \varepsilon/2\big),\end{aligned}$$

where the first equality is due to symmetry and the second equality is because the marginal processes are the same under $\mathbb{P}_0$ and $\mathbb{P}_{t_\star}$. Now we can bound from below the total variation

distance in question by considering the event $\{|f(G_n^1) - f(G_n^2)| \le \varepsilon\}$ under $\mathbb{P}_0$ and $\mathbb{P}_{t_\star}$. For any $n \ge t_\star$ we have that

$$\begin{aligned}
&\mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) \\
&\quad \ge \mathbb{P}_{t_\star}\big(|f(G_n^1) - f(G_n^2)| \le \varepsilon\big) - \mathbb{P}_0\big(|f(G_n^1) - f(G_n^2)| \le \varepsilon\big) \\
&\quad \ge 1 - 2\mathbb{P}_0\big(|f(G_n^1) - f(G_{t_\star}^1)| > \varepsilon/2\big) - \mathbb{P}_0\big(|f(G_n^1) - f(G_n^2)| \le \varepsilon\big).
\end{aligned}$$

Taking limits as $n \to \infty$, we obtain that

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{TV}\big(\mathcal{CG}(n, t_\star, S), \mathcal{G}(n, S)^{\otimes 2}\big) \\
&\qquad \ge 1 - 2\mathbb{P}_0\big(|f_\infty^1 - f(G_{t_\star}^1)| > \varepsilon/2\big) - \mathbb{P}_0\big(|f_\infty^1 - f_\infty^2| \le \varepsilon\big).
\end{aligned}$$

Since $f(G_{t_\star}^1) \to f_\infty^1$ almost surely as $t_\star \to \infty$, we also have that

$$\mathbb{P}_0\big(|f_\infty^1 - f(G_{t_\star}^1)| > \varepsilon/2\big) \to 0$$

as $t_\star \to \infty$. Thus there exists $t' = t'(\delta)$ such that $\mathbb{P}_0(|f_\infty^1 - f(G_{t_\star}^1)| > \varepsilon/2) \le \delta/4$ for every $t_\star \ge t'$. Combining this with (3.2) shows (3.1) and concludes the proof. □

3.2. *Applications to PA and UA trees.* Here we show how Theorem 1.6 can be applied to PA and UA trees, in order to prove Theorems 1.7 and 1.8. In order to apply Theorem 1.6, we have to find a function f such that

$$\lim_{t\to\infty} f(G_t) =: f_\infty$$

exists almost surely and that f_∞ is an absolutely continuous random variable, where $\{G_t\}_{t \ge |S|}$ is a sequence of randomly growing graphs with seed S and attachment rule $\mathcal{G}$, and where $\mathcal{G}$ corresponds to either PA or UA trees.

We first argue that it is enough to show this for the special case when the seed is S_2, the unique tree on two vertices, as this implies the same for any seed tree S on at least two vertices. Indeed, for a tree S on at least two vertices, $\mathrm{PA}(n, S)$ has the same distribution as $\mathrm{PA}(n, S_2)$ conditioned on $\mathrm{PA}(|S|, S_2) = S$ (an event which has positive probability), and therefore, the function f that works for the seed S_2 (i.e., which has the desired properties) also works when the seed is S. The same argument works for UA trees as well. More generally, suppose that $\mathcal{G}$ is a Markov sequential attachment rule and we have a function f satisfying the desired properties when the seed is S'. Then the same function f also satisfies the desired properties whenever the seed S satisfies $S \in \mathrm{Range}(\mathcal{G}, S')$. For PA and UA trees we simply use that $\mathrm{Range}(\mathrm{PA}, S_2) = \mathrm{Range}(\mathrm{UA}, S_2)$ consists of all finite trees on at least two vertices.

Therefore, in the following we may, and thus will, assume that the seed is $S = S_2$. We start with PA trees, for which considering the normalized maximum degree suffices.

PROOF OF THEOREM 1.7. For a graph G, define $f(G) := \Delta(G)/\sqrt{|G|}$, where recall that $\Delta(G)$ is the maximum degree in G. Móri [35], Theorem 3.1, showed that the limit $f_\infty := \lim_{n\to\infty} f(\mathrm{PA}(n, S_2))$ exists almost surely, and moreover that the limit is almost surely positive, finite, and it has an absolutely continuous distribution. Now applying Theorem 1.6 yields the desired conclusion. □

We next present a method that works equally well for both PA and UA trees, with only minor changes needed between the two cases. Accordingly, we present a unified proof for Theorems 1.7 and 1.8, and throughout the proof we will always explain what differs for PA and UA trees. The proof is based on a notion of centrality in trees, which we detail below.

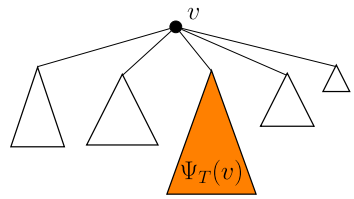


FIG. 3. *The anti-centrality $\Psi_T(v)$ of vertex v in tree T is the size of the largest pendent subtree.*

Given a tree T and a distinguished vertex v in the tree, let (T, v) be the rooted tree with root v. For any other vertex u, $(T, v)_{u\downarrow}$ is the rooted subtree of (T, v) whose root is u and whose vertex set contains all vertices w such that the unique path connecting w and v contains u. The *anti-centrality* of a vertex v in a tree T is defined as

$$\Psi_T(v) := \max_{u \in \mathcal{N}_v(T)} \big|(T, v)_{u\downarrow}\big|,$$

where $\mathcal{N}_v(T) := \{u \in V(T) : (u, v) \in E(T)\}$ is the neighborhood of v in T; see Figure 3 for an illustration. Note that $\Psi_T(v)$ is efficiently computable (i.e., in poly$(|T|)$ time, e.g., using a breadth first search (BFS) algorithm). A *centroid* is a vertex that has minimum anti-centrality. Note that there can be multiple centroids, but only at most two (see, e.g., [27], Lemma 2.1). If there is a unique centroid (which is often the case), then we refer to it as *the* centroid. Properties of this centrality measure and of the corresponding centroid(s) have been widely studied, both for trees in general and also more specifically in a variety of sequentially-generated trees, including PA and UA trees (see, e.g., [27] and the references therein). Centroids and centrality were also used as a key tool in root-finding algorithms in PA and UA trees [8, 17, 32].

In the following $\{T_n\}_{n\geq 2}$ denotes a sequence of trees started from the seed S_2 and grown according to PA or UA. To abbreviate notation, we write $\Psi_n(v) := \Psi_{T_n}(v)$ for a vertex $v \in V(T_n)$. Recall that for a vertex v in the tree T_n, we denote by $\tau(v)$ the *timestamp* of v. That is, $\tau(v) = k$ if v is not in T_{k-1} but is introduced in T_k. In the following when we refer to "a fixed vertex v," we mean that the timestamp $\tau(v)$ of v is fixed (i.e., it does not change with n). The following theorem describes properties of the asymptotic behavior of the anti-centrality of a fixed vertex v in PA and UA trees.

THEOREM 3.1. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed S_2 and grown according to PA or UA. Let v be a fixed vertex. Then the limit*

$$\Psi(v) := \lim_{n\to\infty} \frac{1}{n}\Psi_n(v)$$

exists almost surely. Furthermore, $\Psi(v)$ is an absolutely continuous random variable.

We refer to $\Psi(v)$ as the *limiting anti-centrality* of v. We defer the proof of this theorem to Section 3.3, where, in addition to Theorem 3.1, we also prove a distributional representation of $\Psi(v)$; see Theorem 3.3. The key insight behind the proof is that the evolution of the sizes of the subtrees around v can be described in terms of Pólya urn processes (see [42], Section 4). The limits of these Pólya urn processes are absolutely continuous random variables, from which we can show that $\Psi(v)$ is also an absolutely continuous random variable. The structure of $\Psi(v)$ is the same in both PA and UA trees (with only minor differences in the details), which allows us to develop techniques and proofs that simultaneously work for both models of random trees.

We are particularly interested in the anti-centrality of the centroid(s). Note that even if the tree T has two centroids, the anti-centrality of the two centroids is equal, by definition. If $\theta(T)$ is a centroid of the tree T, then

$$\Psi_T\big(\theta(T)\big) = \min_{v\in T} \max_{u\in\mathcal{N}_v(T)} \big|(T,v)_{u\downarrow}\big|.$$

Turning to the sequence of trees $\{T_n\}_{n\geq 2}$, let $\theta(n) := \theta(T_n)$ denote a centroid of T_n. Jog and Loh proved in [27], for both PA and UA trees, that almost surely the centroid only changes finitely many times. That is, the limit $\theta := \lim_{n\to\infty}\theta(n)$ exists almost surely; we call θ the *limiting centroid* of the sequence of trees $\{T_n\}_{n\geq 2}$. Together with Theorem 3.1 this implies the following corollary.

COROLLARY 3.2. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed S_2 and grown according to PA or UA. Moreover, let $\theta(n) := \theta(T_n)$ denote a centroid of T_n. Then the limit*

$$\lim_{n\to\infty}\frac{1}{n}\Psi_n\big(\theta(n)\big) = \lim_{n\to\infty}\frac{1}{n}\min_{v\in T_n}\max_{u\in\mathcal{N}_v(T_n)}\big|(T_n,v)_{u\downarrow}\big|$$

exists almost surely and is an absolutely continuous random variable.

PROOF. By [27], the centroid stabilizes almost surely, that is, the limiting centroid $\theta := \lim_{n\to\infty}\theta(n)$ exists almost surely. Let $v_1, v_2, \ldots, v_n$ denote the vertices in T_n, labeled in order of appearance; that is, $\tau(v_k) = k$ for $k > 2$ (and v_1 and v_2 are the two vertices in the initial tree T_2). Fix $k \geq 1$ and let $E_k := \{\theta = v_k\}$ be the event that the limiting centroid is v_k. Let $E := \bigcup_{k\geq 1} E_k$ and note that $\mathbb{P}(E) = 1$. On the event E_k we have that $\lim_{n\to\infty}\frac{1}{n}\Psi_n(\theta(n)) = \Psi(v_k)$, so altogether we have that $\lim_{n\to\infty}\frac{1}{n}\Psi_n(\theta(n)) = \Psi(\theta)$ and thus the limit exists almost surely. To see that the limit is absolutely continuous, let F be a set with Lebesgue measure zero. Then

$$\begin{aligned}\mathbb{P}\big(\Psi(\theta)\in F\big) &= \sum_{k\geq 1}\mathbb{P}\big(\{\Psi(\theta)\in F\}\cap E_k\big) = \sum_{k\geq 1}\mathbb{P}\big(\{\Psi(v_k)\in F\}\cap E_k\big)\\ &\leq \sum_{k\geq 1}\mathbb{P}\big(\Psi(v_k)\in F\big) = 0,\end{aligned}$$

where in the second equality we used the definition of E_k and in the last equality we used that $\Psi(v_k)$ is absolutely continuous for any fixed $k \geq 1$. □

Corollary 3.2 directly implies Theorems 1.7 and 1.8, as follows.

PROOF OF THEOREMS 1.7 AND 1.8. For a tree T, define

$$f(T) := \frac{1}{|T|}\min_{v\in T}\max_{u\in\mathcal{N}_v(T)}\big|(T,v)_{u\downarrow}\big|.$$

By Corollary 3.2, the limit $f_\infty := f(T_n)$ exists almost surely and is absolutely continuous, for both PA and UA trees. Now applying Theorem 1.6 yields the desired conclusion. □

3.3. *The distribution of subtree sizes and anti-centrality.* In this section we derive the limiting distribution of the sizes of the subtrees around a fixed vertex v, and using this we derive a distributional representation of the limiting anti-centrality $\Psi(v)$. Theorem 3.1 then follows immediately. Before we state the main theorem of this section, we recall the definition of the timestamp $\tau(v)$ of v: $\tau(v) = k$ if v is not in T_{k-1} but is in T_k. In particular, we use the convention that the timestamp of both vertices in the initial tree T_2 is 2.

THEOREM 3.3. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed S_2 and grown according to PA or UA. Let v be a fixed vertex. Let $\{\varphi_k\}_{k\geq 0}$ be mutually independent random variables, all of them having a beta distribution, with parameters as follows*:

$$\varphi_0 \sim \begin{cases} \mathrm{Beta}(\tau(v)-1,1) & \text{for UA,} \\ \mathrm{Beta}\left(\tau(v)-\frac{3}{2},\frac{1}{2}\right) & \text{for PA,} \end{cases}$$

and for $k \geq 1$, let

$$\varphi_k \sim \begin{cases} \mathrm{Beta}(1,1) & \text{for UA,} \\ \mathrm{Beta}\left(\frac{1}{2},\frac{k+1}{2}\right) & \text{for PA.} \end{cases}$$

We then define the random variables $\{\psi_\ell\}_{\ell\geq 0}$ as follows: $\psi_0 := \varphi_0$, *and for $\ell \geq 1$ let*

$$\psi_\ell := \varphi_\ell \prod_{i=0}^{\ell-1}(1-\varphi_i).$$

The limiting anti-centrality $\Psi(v)$ of v exists almost surely and has the following distributional representation:

$$\Psi(v) \stackrel{d}{=} \max_{\ell\geq 0} \psi_\ell. \tag{3.3}$$

In this representation ψ_ℓ is the asymptotic normalized size of the ℓth subtree around v; here counting starts at $\ell = 0$ and subtrees are ordered according to their first appearance around v. Similar representations—of various limiting quantities using a sequence of independent (beta) random variables—are common in the study of preferential attachment, uniform attachment, and related random graph models (see, e.g., [5, 42]). We refer the reader to [42], Section 4, for an exposition.

PROOF. We first prove the claim for UA trees. Note that v is a leaf in $T_{\tau(v)}$. Let u_0 denote the neighbor of v in $T_{\tau(v)}$ and let e_0 denote the edge connecting v and u_0. For $n \geq \tau(v)$, the edge e_0 partitions T_n into two subtrees: $(T_n, v)_{u_0\downarrow}$ and $T_n \setminus (T_n, v)_{u_0\downarrow}$. When a new vertex joins the tree, it attaches to an existing vertex uniformly at random. Therefore, the probability of the new vertex joining either one of these two subtrees is proportional to their size. Thus the evolution of the pair of subtree sizes, $(|(T_n, v)_{u_0\downarrow}|, n - |(T_n, v)_{u_0\downarrow}|)$, follows a classical Pólya urn. Initially, at time $n = \tau(v)$, the pair of subtree sizes is $(\tau(v)-1, 1)$. Therefore, by classical results on Pólya urns (see, e.g., [42], Section 4.5 and Example 4.7), the limit

$$\varphi_0 := \lim_{n\to\infty} \frac{1}{n}|(T_n, v)_{u_0\downarrow}| \tag{3.4}$$

exists almost surely and $\varphi_0 \sim \mathrm{Beta}(\tau(v)-1, 1)$.

Next, let u_1 denote the first vertex that attaches to v with $\tau(u_1) > \tau(v)$, and let e_1 denote the edge connecting v and u_1. (Note that almost surely $\tau(u_1) < \infty$.) For $n \geq \tau(u_1)$, the edges e_0 and e_1 partition the tree T_n into three subtrees: $(T_n, v)_{u_0\downarrow}$, $(T_n, v)_{u_1\downarrow}$, and $T_n \setminus ((T_n, v)_{u_0\downarrow} \cup (T_n, v)_{u_1\downarrow})$. When a new vertex joins the tree, it attaches to an existing vertex uniformly at random. We can view this as a multi-stage process as follows. First, the vertex decides whether it will join the subtree $(T_n, v)_{u_0\downarrow}$ or the subtree $T_n \setminus (T_n, v)_{u_0\downarrow}$; it does so by flipping a coin, with the probability of choosing either option being proportional to the size of the respective subtree. Next, if the vertex decides to join the subtree $T_n \setminus (T_n, v)_{u_0\downarrow}$, it then chooses whether to join the subtree $(T_n, v)_{u_1\downarrow}$ or the subtree $T_n \setminus ((T_n, v)_{u_0\downarrow} \cup (T_n, v)_{u_1\downarrow})$;

it again does so by flipping a coin, with the probability of choosing either option being proportional to the size of the respective subtree. This second coin flip is independent of the first coin flip. Finally, once the vertex has decided which of the three subtrees to join, it attaches to a vertex chosen uniformly at random from the given subtree.

From this construction it is immediate that, when viewed at the times when the new vertex joins the subtree $T_n \setminus (T_n, v)_{u_0\downarrow}$, the pair

$$\big(|(T_n, v)_{u_1\downarrow}|, n - |(T_n, v)_{u_0\downarrow}| - |(T_n, v)_{u_1\downarrow}|\big) \tag{3.5}$$

evolves as a classical Pólya urn started from $(1, 1)$. Thus the limit

$$\varphi_1 := \lim_{n\to\infty} \frac{|(T_n, v)_{u_1\downarrow}|}{n - |(T_n, v)_{u_0\downarrow}|} \tag{3.6}$$

exists almost surely and $\varphi_1 \sim \mathrm{Beta}(1, 1)$ (in other words, φ_1 is uniform on the interval $[0, 1]$). Moreover, the evolution of the Pólya urn describing the pair in (3.5) is independent of the process that determines the times at which the subtree $T_n \setminus (T_n, v)_{u_0\downarrow}$ increases, which means that φ_1 and φ_0 are independent. Putting together (3.4) and (3.6), we obtain that

$$\lim_{n\to\infty} \frac{1}{n}|(T_n, v)_{u_1\downarrow}| = \lim_{n\to\infty} \frac{|(T_n, v)_{u_1\downarrow}|}{n - |(T_n, v)_{u_0\downarrow}|} \frac{n - |(T_n, v)_{u_0\downarrow}|}{n} = \varphi_1(1 - \varphi_0) = \psi_1$$

almost surely.

We can then iterate this argument. For $\ell \geq 2$, let u_ℓ denote the ℓth vertex to attach to v. The random variables $\varphi_2, \varphi_3, \ldots$ can be defined inductively by the limit

$$\varphi_\ell := \lim_{n\to\infty} \frac{|(T_n, v)_{u_\ell\downarrow}|}{n - \sum_{i=0}^{\ell-1} |(T_n, v)_{u_i\downarrow}|};$$

the same argument as above shows that this limit exists almost surely, $\varphi_\ell \sim \mathrm{Beta}(1, 1)$ for every $\ell \geq 1$, and that φ_ℓ is independent of $\varphi_0, \varphi_1, \ldots, \varphi_{\ell-1}$. Subsequently, this implies by induction that

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}|(T_n, v)_{u_\ell\downarrow}| &= \lim_{n\to\infty} \frac{|(T_n, v)_{u_\ell\downarrow}|}{n - \sum_{i=0}^{\ell-1} |(T_n, v)_{u_i\downarrow}|} \frac{n - \sum_{i=0}^{\ell-1} |(T_n, v)_{u_i\downarrow}|}{n} \\ &= \varphi_\ell \prod_{i=0}^{\ell-1} (1 - \varphi_i) = \psi_\ell\end{aligned}$$

almost surely. We have thus shown that the asymptotic normalized size of the ℓth subtree around v is given by ψ_ℓ. What remains is to understand how the subtree sizes of these fixed neighbors of v relate to the anti-centrality of v.

Define the event $E_k := \{\varphi_k > 1/2\}$ and let $E := \bigcup_{k\geq 1} E_k$. The events $\{E_k\}_{k\geq 1}$ are mutually independent and $\mathbb{P}(E_k) = 1/2$ for every $k \geq 1$. Therefore, $\mathbb{P}(E) = 1$. Since E_k holds if and only if $\varphi_k > 1 - \varphi_k$, the event E_k is equivalent to the event that

$$\lim_{n\to\infty} \frac{1}{n}|(T_n, v)_{u_k\downarrow}| > \lim_{n\to\infty} \frac{1}{n}\left(n - \sum_{\ell=0}^{k} |(T_n, v)_{u_\ell\downarrow}|\right)$$

holds. Thus on the event E_k we have, for all n large enough, that

$$|(T_n, v)_{u_k\downarrow}| > n - \sum_{\ell=0}^{k} |(T_n, v)_{u_\ell\downarrow}|.$$

Consider the neighbors of v, and note that together the pendant subtrees make up all of the tree (except for v). Thus for any $u \in \mathcal{N}_v(T_n) \setminus \{u_0, u_1, \dots, u_k\}$ we have that

$$\big|(T_n, v)_{u\downarrow}\big| \le \sum_{w \in \mathcal{N}_v(T_n)\setminus\{u_0,u_1,\dots,u_k\}} \big|(T_n, v)_{w\downarrow}\big| = n - 1 - \sum_{\ell=0}^{k} \big|(T_n, v)_{u_\ell\downarrow}\big|.$$

Therefore, it follows that, on the event E_k, we have that

$$\Psi_n(v) = \max_{\ell\in\{0,1,\dots,k\}} \big|(T_n, v)_{u_\ell\downarrow}\big|$$

for all n large enough. Thus dividing by n and taking limits, we have that, on the event E_k, the limit $\Psi(v) := \lim_{n\to\infty} \frac{1}{n}\Psi_n(v)$ exists and moreover

$$\Psi(v) = \max_{\ell\in\{0,1,\dots,k\}} \psi_\ell.$$

Consequently, on the event E, the limit $\Psi(v) := \lim_{n\to\infty} \frac{1}{n}\Psi_n(v)$ exists and moreover $\Psi(v) = \max_{\ell\ge 0} \psi_\ell$. Since E holds almost surely, this concludes the proof of (3.3) for UA trees.

For PA trees the arguments are similar, so we only explain the differences. In PA, when a new vertex joins the tree, it attaches to an existing vertex with probability proportional to its degree. Thus, if we partition the tree into finitely many subtrees, the probability that the new vertex joins a particular subtree is proportional to the sum of the degrees of the vertices in the subtree. Moreover, when a vertex joins a particular subtree, it increases the sum of the degrees in the subtree by 2, due to the new edge. For more details, see [42], Section 4.5 and Example 4.11.

Thus there are two differences in the analysis of subtrees above: (1) the quantity associated with a subtree that we analyze is now the sum of the degrees of the vertices in the subtree (instead of the number of vertices in the subtree), and (2) the Pólya urns that arise have replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ (see [42], Section 4.5). The first change also means that the initial conditions of the appropriate Pólya urns are different. Specifically, the limiting random variable φ_0 arises from a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and initial condition $(2\tau(v) - 3, 1)$, which is why $\varphi_0 \sim \mathrm{Beta}(\tau(v) - \frac{3}{2}, \frac{1}{2})$. For $k \ge 1$, the limiting random variable φ_k arises from a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and initial condition $(1, k + 1)$, which is why $\varphi_k \sim \mathrm{Beta}(\frac{1}{2}, \frac{k+1}{2})$.

There is one more subtle point here: we are interested in the asymptotic behavior of the sizes of various subtrees (that is, the number of vertices in the subtrees), but the analysis concerns the sum of the degrees of the vertices in the subtrees. However, the map $x \mapsto 2x - 1$ takes the number of vertices in a subtree to the sum of the degrees of the vertices in the subtree (this uses the fact that we are considering subtrees where there is exactly one edge exiting the subtree). The normalization factor also differs by essentially a factor of 2: it is n when the considering the number of vertices and $2n - 2$ when considering the sum of the degrees. Thus after normalization the quantity that we care about (subtree size) is asymptotically the same as the quantity that we analyze (sum of the degrees in a subtree).

With these changes we have thus determined that the asymptotic normalized size of the ℓth subtree around v is given by ψ_ℓ for PA trees. What remains is to show (3.3) for PA trees. Since the random variables $\{\varphi_k\}_{k\ge 1}$ are no longer i.i.d. uniform on $[0, 1]$ (as in the case of UA trees), a different argument is needed here. For $k \ge 1$ define the event

$$E_k := \left\{ \max_{\ell\in\{0,1,\dots,k\}} \psi_\ell > 1 - \sum_{\ell=0}^{k} \psi_\ell \right\}$$

and let $E := \bigcup_{k\geq 0} E_k$. An analogous argument as above shows that on the event E_k we have that $\Psi(v) := \lim_{n\to\infty} \frac{1}{n}\Psi_n(v)$ exists and moreover $\Psi(v) = \max_{\ell\in\{0,1,\ldots,k\}} \psi_\ell$. Thus on the event E we have that $\Psi(v) := \lim_{n\to\infty} \frac{1}{n}\Psi_n(v)$ exists and moreover $\Psi(v) = \max_{\ell\geq 0} \psi_\ell$. What remains to show is that $\mathbb{P}(E) = 1$, which is equivalent to showing that $\lim_{k\to\infty} \mathbb{P}(E_k) = 1$, since $\{E_k\}_{k\geq 0}$ is an increasing sequence of events. This, in turn, follows from the fact that $1 - \sum_{\ell=0}^{k} \psi_\ell \to 0$ in probability as $k \to \infty$. To see that this convergence in probability holds, first observe that $1 - \sum_{\ell=0}^{k} \psi_\ell = \prod_{\ell=0}^{k}(1-\varphi_k)$. Then by independence we have that

$$\mathbb{E}\left[1 - \sum_{\ell=0}^{k} \psi_\ell\right] = \prod_{\ell=0}^{k} \mathbb{E}[1-\varphi_k] = \frac{1}{2\tau(v)-2} \prod_{\ell=1}^{k} \frac{\ell+1}{\ell+2} = \frac{1}{(\tau(v)-1)(k+2)},$$

which goes to 0 as $k \to \infty$. The conclusion then follows from Markov's inequality. □

Theorem 3.3 directly implies Theorem 3.1, as we now show.

PROOF OF THEOREM 3.1. By Theorem 3.3, the limiting anti-centrality $\Psi(v)$ exists almost surely. Moreover, it satisfies the distributional representation given in (3.3). That is, it is the maximum of countably many absolutely continuous random variables. As such, it is absolutely continuous as well. Indeed, if F is a set with Lebesgue measure zero, then

$$\mathbb{P}\big(\Psi(v) \in F\big) = \mathbb{P}\Big(\max_{\ell\geq 0} \psi_\ell \in F\Big) \leq \sum_{\ell\geq 0} \mathbb{P}(\psi_\ell \in F) = 0.$$

□

4. An initial, coarse estimate of $t_\star$. We now turn to the problem of estimating $t_\star$. The estimator that we use to prove Theorem 1.9 is somewhat involved, so in this section we first study a simpler estimator. The guarantees we prove for this simpler estimator are weaker than those in Theorem 1.9 (see Theorem 4.1 below), but studying this simpler estimator highlights some of the key ideas that also go into the more involved estimator studied subsequently in Section 5. Moreover, as we shall see in Section 5, our estimator for $t_\star$ that achieves vanishing relative error needs as input an initial, coarse estimate of $t_\star$—and the simple estimator studied in this section provides this.

In this section we will thus prove the following result.

THEOREM 4.1 (A coarse estimate of $t_\star$ in PA and UA trees). *Let $S = S_2$ be the unique tree on two vertices and let $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. There exists an estimator $\widehat{t}_n \equiv \widehat{t}(T_n^1, T_n^2)$, computable in polynomial time, such that*

$$\lim_{t_\star\to\infty} \liminf_{n\to\infty} \mathbb{P}\left(\frac{t_\star}{\log t_\star} \leq \widehat{t}_n \leq t_\star \log t_\star\right) = 1.$$

The same result also holds when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

We now describe the estimator used to prove Theorem 4.1. Recall all the notation introduced in Sections 3.2 and 3.3, which we will use here. Moreover, for anything introduced previously in these sections, if we add a superscript i to it (where $i \in \{1, 2\}$), this means that it is the appropriate object in the tree T_n^i. For instance, $\theta^1(n)$ and $\theta^2(n)$ are the centroids in T_n^1 and T_n^2, respectively.

The main idea is to consider the minimum anti-centrality in the two trees T_n^1 and T_n^2. In other words, we consider the sizes of the largest pendent subtrees of the two centroids. The heuristic, which we will make precise, is as follows. If $t_\star$ is large, then the centroids in T_n^1 and T_n^2 correspond to the same vertex, with probability close to 1. If this is the case, then the sizes of the largest pendent subtrees of the centroids should be similar, and their difference

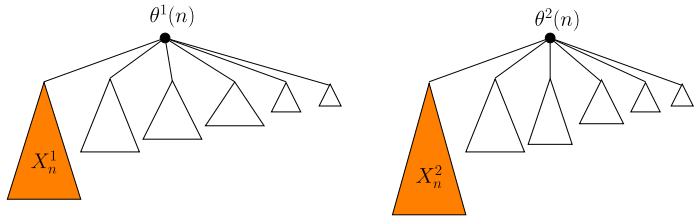


FIG. 4. *The pendent subtrees of the centroids $\theta^1(n)$ and $\theta^2(n)$ in T_n^1 and T_n^2, respectively, are ordered in decreasing order. The estimator studied in Section 4 is a function of the (normalized) sizes of the largest pendent subtrees, X_n^1 and X_n^2.*

should concentrate on some function of n and $t_\star$—which should be a function of only $t_\star$ in the limit as $n \to \infty$. Estimating this function and inverting it then allows us to estimate $t_\star$. See Figure 4 for an illustration.

Thus we define, for $i \in \{1, 2\}$, the random variable

$$X_n^i := \frac{1}{n} \min_{v \in T_n^i} \max_{u \in \mathcal{N}_v(T_n^i)} \left|(T_n^i, v)_{u\downarrow}\right| = \frac{1}{n} \Psi_{T_n^i}(\theta^i(n)). \tag{4.1}$$

Now define

$$Y_n := \frac{(X_n^1 - X_n^2)^2}{2X_n^1(1 - X_n^1)}. \tag{4.2}$$

As we shall see, Y_n is concentrated around $1/t_\star$, so we can define the estimator $\widehat{t}_n := 1/Y_n$. Theorem 4.1 then follows immediately from the following result.

THEOREM 4.2. *Let $S = S_2$ be the unique tree on two vertices and let $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. Define Y_n via (4.1) and (4.2). We have that*

$$\lim_{t_\star \to \infty} \liminf_{n \to \infty} \mathbb{P}\left(\frac{1}{t_\star \log t_\star} \le Y_n \le \frac{\log t_\star}{t_\star} \right) = 1.$$

The same result also holds when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

In the remainder of this section we prove this theorem. We start in Section 4.1 with some preliminaries: specifically, we define a couple of "nice" events on the space of sequences of growing trees, on which we will obtain bounds for Y_n. We prove a first moment estimate for Y_n in Section 4.2. We then prove Theorem 4.2 in Section 4.3, using the fact that the previously defined "nice" events have probability close to 1. We defer the proof of this latter fact to Appendix A.

4.1. *Preliminaries.* We start by introducing some notation on labeling vertices. Let $\{T_n\}_{n \ge 2}$ be a growing sequence of trees started from the seed $S = S_2$, where at each step we add a single new node and a new edge. We denote the vertices of T_n by $v_1, v_2, \dots, v_n$, where v_1 and v_2 are the two initial vertices in S, and for $k \ge 3$, v_k is the unique vertex with timestamp k. As before, we write $\Psi_n(v) := \Psi_{T_n}(v)$ for a vertex $v \in V(T_n)$. We write $\widetilde{v}_{i,n}(1)$ for the neighbor of v_i that is the root of the largest subtree of (T_n, v_i) (assuming that there is a unique largest subtree). With this notation we have that $\Psi_n(v_i) = |(T_n, v_i)_{\widetilde{v}_{i,n}(1)\downarrow}|$. More generally, for any $k \ge 1$ we write $\widetilde{v}_{i,n}(k)$ for the neighbor of v_i that is the root of the kth largest subtree of (T_n, v_i) (assuming that there is a unique such vertex). Finally, we write $\widetilde{\theta}_n(1)$ for the neighbor of the centroid $\theta(n)$ that is the root of the largest subtree of $(T_n, \theta(n))$ (assuming that the centroid is unique and that there is a unique largest subtree).

We are now ready to define what we mean by the "nice" event on the space of sequences of growing trees.

DEFINITION 4.3 (The event $\mathcal{A}$). Given a sequence of trees $\{T_n\}_{n\geq t_\star}$, we say that the event $\mathcal{A}$ holds if and only if the following three properties all hold:

(A1) The centroid $\theta(n)$ is unique for all $n \geq t_\star$ and $\theta(n) = \theta(t_\star)$ for all $n \geq t_\star$.
(A2) The vertex $\widetilde{\theta}_n(1)$ is uniquely defined for all $n \geq t_\star$ and $\widetilde{\theta}_n(1) = \widetilde{\theta}_{t_\star}(1)$ for all $n \geq t_\star$.
(A3) For all $n \geq t_\star$ we have that

$$\left|\frac{1}{n}\Psi_n\big(\theta(n)\big) - \frac{1}{t_\star}\Psi_{t_\star}\big(\theta(t_\star)\big)\right| \leq \frac{1}{t_\star^{1/3}} \min\left\{\frac{1}{t_\star}\Psi_{t_\star}\big(\theta(t_\star)\big), 1 - \frac{1}{t_\star}\Psi_{t_\star}\big(\theta(t_\star)\big)\right\}. \tag{4.3}$$

The exponent $1/3$ in (4.3) is chosen for simplicity; any positive constant that is less than $1/2$ is a good choice for everything that follows. Furthermore, we always have that $\Psi_{t_\star}(\theta(t_\star)) \leq t_\star/2$—this is a known property of tree centroids (see, e.g., [27], Lemma 2.1)—so the minimum in (4.3) is always attained by the first term; we include the second term in the definition just for clarity. Given a sequence of trees $\{T_n\}_{n\geq 2}$, we say that the event $\mathcal{A}$ holds if and only if it holds for the subsequence $\{T_n\}_{n\geq t_\star}$. The event $\mathcal{A}$ clearly depends on $t_\star$, but we choose to omit $t_\star$ from the notation in order to keep notation lighter. The following lemma shows that for PA and UA trees the event $\mathcal{A}$ holds with probability close to 1 when $t_\star$ is large.

LEMMA 4.4. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed S and grown according to PA or UA. There exists a finite constant C such that for every $t_\star \geq 2$ we have that*

$$\mathbb{P}(\mathcal{A}^c) \leq \frac{C}{\log t_\star}, \tag{4.4}$$

where $\mathcal{A}^c$ denotes the complement of $\mathcal{A}$.

The proof of Lemma 4.4 is deferred to Appendix A.

The intuition behind defining $\mathcal{A}$ in this way is as follows. On the event $\mathcal{A}$, both the centroid and the largest subtree of the centroid do not change locations within the tree for $n \geq t_\star$. Hence, by conditioning on the tree at time $t_\star$, studying $\Psi_n(\theta(n))$ essentially amounts to understanding the growth of a *fixed* subtree that is present in $T_{t_\star}$. Since the sizes of fixed subtrees grow according to Pólya urn processes (in PA and UA trees), their distributions are very well understood.

We are interested in a pair of correlated randomly growing (either PA or UA) trees $\{(T_n^1, T_n^2)\}_{n\geq 2}$. Let $\mathcal{A}^1$ and $\mathcal{A}^2$ denote the "nice" events corresponding to $\{T_n^1\}_{n\geq 2}$ and $\{T_n^2\}_{n\geq 2}$. Since $T_n^1 = T_n^2$ for all $n \leq t_\star$, we have, in particular, that $\theta^1(t_\star) = \theta^2(t_\star) =: \theta(t_\star)$ and also that $\widetilde{\theta}_{t_\star}^1(1) = \widetilde{\theta}_{t_\star}^2(1) =: \widetilde{\theta}_{t_\star}(1)$. A key observation is that on the event $\mathcal{A}^1 \cap \mathcal{A}^2$ we have that $\theta^1(n) = \theta^2(n) = \theta(t_\star)$ for all $n \geq t_\star$ and that $\widetilde{\theta}_n^1(1) = \widetilde{\theta}_n^2(1) = \widetilde{\theta}_{t_\star}(1)$ for all $n \geq t_\star$, which implies that on the event $\mathcal{A}^1 \cap \mathcal{A}^2$ we have that $X_n^i = \frac{1}{n}|(T_n^i, \theta(t_\star))_{\widetilde{\theta}_{t_\star}(1)\downarrow}|$ for $i \in \{1, 2\}$ and all $n \geq t_\star$. Thus in order to understand the behavior of the statistic Y_n on the event $\mathcal{A}^1 \cap \mathcal{A}^2$, it suffices to condition on the tree at time $t_\star$ and then analyze the behavior of fixed subtrees. We do this next.

Condition now on the tree $T_{t_\star}^1 = T_{t_\star}^2 =: T_{t_\star}$; that is, assume that $T_{t_\star}$ is given. To abbreviate notation, we write $\theta := \theta(t_\star)$ and $\widetilde{\theta}(1) := \widetilde{\theta}_{t_\star}(1)$; importantly, note that these are now *fixed* vertices (i.e., they do not change with n). Define the random variables

$$Z_n^i := \frac{1}{n}|(T_n^i, \theta)_{\widetilde{\theta}(1)\downarrow}|$$

for $i \in \{1, 2\}$ and $n \geq t_\star$. As observed above, on the event $\mathcal{A}^i$ we have that $X_n^i = Z_n^i$ for $n \geq t_\star$.

In UA trees, the evolution of $(nZ_n^i, n - nZ_n^i)$ for $n \geq t_\star$ follows a classical Pólya urn with initial condition $(\Psi_{t_\star}(\theta), t_\star - \Psi_{t_\star}(\theta))$, for $i \in \{1, 2\}$. Moreover, the Pólya urns for $i = 1$ and $i = 2$ are independent (recall that we are conditioning on $T_{t_\star}$, so this is conditional independence given $T_{t_\star}$). In PA trees, the evolution of $(2nZ_n^i - 1, 2n - 2nZ_n^i - 1)$ for $n \geq t_\star$ follows a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and initial condition $(2\Psi_{t_\star}(\theta) - 1, 2t_\star - 2\Psi_{t_\star}(\theta) - 1)$, for $i \in \{1, 2\}$. Moreover, the Pólya urns for $i = 1$ and $i = 2$ are independent (again, this is conditional independence given $T_{t_\star}$).

Thus by classical results on Pólya urns it follows that the limiting random variables

$$Z^i := \lim_{n \to \infty} Z_n^i$$

exist almost surely for $i \in \{1, 2\}$, for both PA and UA trees. Moreover, Z^1 and Z^2 are i.i.d. (again, this is conditional independence given $T_{t_\star}$) beta random variables, with parameters given as follows:

$$Z \sim \begin{cases} \mathrm{Beta}\big(\Psi_{t_\star}(\theta), t_\star - \Psi_{t_\star}(\theta)\big) & \text{for UA,} \\ \mathrm{Beta}\Big(\Psi_{t_\star}(\theta) - \dfrac{1}{2}, t_\star - \Psi_{t_\star}(\theta) - \dfrac{1}{2}\Big) & \text{for PA.} \end{cases} \tag{4.5}$$

Here Z is a random variable with the same distribution as Z^1 and Z^2.

From (4.5) it is clear that the quantity $\Psi_{t_\star}(\theta)$ plays an important role in the distribution of Z. We always have that $\Psi_{t_\star}(\theta) \leq t_\star/2$. Typically $\Psi_{t_\star}(\theta)$ is on the order $t_\star$, but with some small probability it can be of smaller order. The following definition and lemma quantify this.

DEFINITION 4.5 (The event $\mathcal{B}$). Let $\mathcal{B}$ denote the following event:

$$\mathcal{B} := \left\{ \frac{t_\star}{\sqrt{\log t_\star}} \leq \Psi_{t_\star}\big(\theta(t_\star)\big) \leq \frac{t_\star}{2} \right\}.$$

The event $\mathcal{B}$ clearly depends on $t_\star$, but we choose to omit $t_\star$ from the notation in order to keep notation lighter. Also, as mentioned above, the bound $\Psi_{t_\star}(\theta) \leq t_\star/2$ always holds, but we still include it in the definition of $\mathcal{B}$ just for clarity.

LEMMA 4.6. *Let $\{T_n\}_{n \geq 2}$ be a sequence of trees started from the seed S and grown according to PA or UA. There exists a finite constant C such that for every $t_\star \geq 2$ we have that*

$$\mathbb{P}(\mathcal{B}^c) \leq \frac{C}{\log^{1/4}(t_\star)}, \tag{4.6}$$

where $\mathcal{B}^c$ denotes the complement of $\mathcal{B}$.

The bound in (4.6) can be improved to $C/\sqrt{\log t_\star}$ for UA trees, but we choose to have a unified theorem for PA and UA trees for simplicity. The proof of Lemma 4.6 is deferred to Appendix A.

4.2. *First moment estimate.* In this subsection we prove the following first moment estimate.

LEMMA 4.7. *Let $(T_n^1, T_n^1) \sim \mathrm{CPA}(n, t_\star, S)$. For all $t_\star$ large enough we have that*

$$\limsup_{n \to \infty} \mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2}] \leq \frac{1 + 3t_\star^{-1/3}}{t_\star}. \tag{4.7}$$

The same bound holds also when $(T_n^1, T_n^1) \sim \mathrm{CUA}(n, t_\star, S)$.

We note that a matching lower bound (of the form $(1-o(1))/t_\star$ as $t_\star \to \infty$) also holds, but since we will not use that direction, we do not give details here.

PROOF. We condition on the tree $T_{t_\star}$ at time $t_\star$; by the tower rule we have that

$$\mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2}] = \mathbb{E}\big[\mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \mid T_{t_\star}]\big]. \tag{4.8}$$

Now given $T_{t_\star}$, observe that property (A3) in Definition 4.3 implies that on the event $\mathcal{A}^1 \cap \mathcal{A}^2$ we have that

$$X_n^1(1-X_n^1) \geq \frac{1}{t_\star}\Psi_{t_\star}(\theta)\bigg(1-\frac{1}{t_\star}\Psi_{t_\star}(\theta)\bigg)(1-t_\star^{-1/3})^2$$

for $n \geq t_\star$. Plugging this inequality into the definition of Y_n we obtain that

$$\begin{aligned}\mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \mid T_{t_\star}] &\leq \frac{\mathbb{E}[(X_n^1 - X_n^2)^2 \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \mid T_{t_\star}]}{2 \cdot \frac{1}{t_\star}\Psi_{t_\star}(\theta)(1-\frac{1}{t_\star}\Psi_{t_\star}(\theta))(1-t_\star^{-1/3})^2} \\ &\leq \frac{\mathbb{E}[(Z_n^1 - Z_n^2)^2 \mid T_{t_\star}]}{2 \cdot \frac{1}{t_\star}\Psi_{t_\star}(\theta)(1-\frac{1}{t_\star}\Psi_{t_\star}(\theta))(1-t_\star^{-1/3})^2},\end{aligned}$$

where the second inequality follows by observing that on the event $\mathcal{A}^1 \cap \mathcal{A}^2$ we have that $X_n^i = Z_n^i$ for $i \in \{1,2\}$, and then removing the indicator to get an upper bound. Taking the limit as $n \to \infty$ and applying the bounded convergence theorem we obtain that

$$\limsup_{n\to\infty} \mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \mid T_{t_\star}] \leq \frac{\mathbb{E}[(Z^1 - Z^2)^2 \mid T_{t_\star}]}{2 \cdot \frac{1}{t_\star}\Psi_{t_\star}(\theta)(1-\frac{1}{t_\star}\Psi_{t_\star}(\theta))(1-t_\star^{-1/3})^2}. \tag{4.9}$$

Now using conditional independence, the limiting conditional distribution obtained in (4.5), and plugging in the variance of the beta distribution, we have that

$$\mathbb{E}\big[(Z^1 - Z^2)^2 \mid T_{t_\star}\big] = 2\operatorname{Var}(Z \mid T_{t_\star}) = \begin{cases} \dfrac{2\Psi_{t_\star}(\theta)(t_\star - \Psi_{t_\star}(\theta))}{t_\star^2(t_\star+1)} & \text{for UA,} \\ \dfrac{2(\Psi_{t_\star}(\theta) - 1/2)(t_\star - \Psi_{t_\star}(\theta) - 1/2)}{(t_\star-1)^2 t_\star} & \text{for PA.} \end{cases}$$

Plugging these formulas into (4.9), we obtain, for both PA and UA trees, that

$$\limsup_{n\to\infty} \mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \mid T_{t_\star}] \leq \bigg(1+\frac{1}{t_\star - 1}\bigg)^2 (1-t_\star^{-1/3})^{-2}\frac{1}{t_\star} \leq \frac{1+3t_\star^{-1/3}}{t_\star},$$

where the second inequality holds for all $t_\star$ large enough. Since this holds for any tree $T_{t_\star}$, taking an expectation and using (4.8) we arrive at (4.7). □

4.3. *Putting everything together*: *Proof of Theorem* 4.2. PROOF OF THEOREM 4.2. We start with the upper bound, which is a consequence of Lemma 4.7 and Markov's inequality. First, by a union bound we have that

$$\mathbb{P}\bigg(Y_n \geq \frac{\log t_\star}{t_\star}\bigg) \leq \mathbb{P}\big((\mathcal{A}^1 \cap \mathcal{A}^2)^c\big) + \mathbb{P}\bigg(\bigg\{Y_n \geq \frac{\log t_\star}{t_\star}\bigg\} \cap \mathcal{A}^1 \cap \mathcal{A}^2\bigg).$$

By a union bound and Lemma 4.4 we have that the first term is at most $C/\log t_\star$ for some constant C, and so it remains to deal with the second term. By Markov's inequality we have that

$$\mathbb{P}\bigg(\bigg\{Y_n \geq \frac{\log t_\star}{t_\star}\bigg\} \cap \mathcal{A}^1 \cap \mathcal{A}^2\bigg) \leq \mathbb{P}\bigg(Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2} \geq \frac{\log t_\star}{t_\star}\bigg) \leq \frac{t_\star}{\log t_\star}\mathbb{E}[Y_n \mathbf{1}_{\mathcal{A}^1 \cap \mathcal{A}^2}].$$

By (4.7) we thus have that

$$\limsup_{n\to\infty} \mathbb{P}\left(Y_n \geq \frac{\log t_\star}{t_\star}\right) \leq \frac{C+2}{\log t_\star}$$

for all $t_\star$ large enough. This expression goes to zero as $t_\star \to \infty$, which concludes the proof of the upper bound.

We now turn to the lower bound. To abbreviate notation, we introduce the quantity $\delta_{t_\star} := (t_\star \log t_\star)^{-1/2}$. Our goal is to show that

$$\lim_{t_\star\to\infty} \limsup_{n\to\infty} \mathbb{P}(Y_n \leq \delta_{t_\star}^2) = 0.$$

Since $Y_n \leq \delta_{t_\star}^2$ implies that $|X_n^1 - X_n^2| \leq \delta_{t_\star}$, we have that

$$\mathbb{P}(Y_n \leq \delta_{t_\star}^2) \leq \mathbb{P}(|X_n^1 - X_n^2| \leq \delta_{t_\star}).$$

By a union bound we have that

$$\begin{aligned}&\mathbb{P}(|X_n^1 - X_n^2| \leq \delta_{t_\star})\\ &\quad\leq \mathbb{P}((\mathcal{A}^1)^c) + \mathbb{P}((\mathcal{A}^2)^c) + \mathbb{P}(\mathcal{B}^c) + \mathbb{P}(\{|X_n^1 - X_n^2| \leq \delta_{t_\star}\} \cap \mathcal{A}^1 \cap \mathcal{A}^2 \cap \mathcal{B}).\end{aligned}$$

By Lemmas 4.4 and 4.6, there exists a finite constant C such that the first three terms in the display above are bounded above by $C/\log^{1/4}(t_\star)$. Since this goes to zero as $t_\star \to \infty$, what remains is to bound the last term in the display above. To do this, we first condition on the tree $T_{t_\star}$. By the tower rule, using also the fact that the event $\mathcal{B}$ is measurable with respect to $T_{t_\star}$, we have that

$$\mathbb{P}(\{|X_n^1 - X_n^2| \leq \delta_{t_\star}\} \cap \mathcal{A}^1 \cap \mathcal{A}^2 \cap \mathcal{B}) = \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{|X_n^1 - X_n^2| \leq \delta_{t_\star}\}} \mathbf{1}_{\mathcal{A}^1\cap\mathcal{A}^2} \mid T_{t_\star}] \mathbf{1}_{\mathcal{B}}]. \tag{4.10}$$

We now fix $T_{t_\star}$ and study the conditional expectation $\mathbb{E}[\mathbf{1}_{\{|X_n^1 - X_n^2| \leq \delta_{t_\star}\}} \mathbf{1}_{\mathcal{A}^1\cap\mathcal{A}^2} \mid T_{t_\star}]$. Recall that on the event $\mathcal{A}^1 \cap \mathcal{A}^2$ we have that $X_n^i = Z_n^i$ for $i \in \{1, 2\}$ and $n \geq t_\star$. Therefore, by the bounded convergence theorem we have that

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}[\mathbf{1}_{\{|X_n^1 - X_n^2| \leq \delta_{t_\star}\}} \mathbf{1}_{\mathcal{A}^1\cap\mathcal{A}^2} \mid T_{t_\star}] &= \mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \leq \delta_{t_\star}\}} \mathbf{1}_{\mathcal{A}^1\cap\mathcal{A}^2} \mid T_{t_\star}]\\ &\leq \mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \leq \delta_{t_\star}\}} \mid T_{t_\star}],\end{aligned} \tag{4.11}$$

where the inequality follows by dropping the second indicator. For notational convenience, and in order to treat the cases of PA and UA trees simultaneously, we introduce

$$(a, b) := \begin{cases} (\Psi_{t_\star}(\theta), t_\star - \Psi_{t_\star}(\theta)) & \text{for UA,}\\ \left(\Psi_{t_\star}(\theta) - \frac{1}{2}, t_\star - \Psi_{t_\star}(\theta) - \frac{1}{2}\right) & \text{for PA.}\end{cases} \tag{4.12}$$

Recall from (4.5) that, conditioned on $T_{t_\star}$, the random variables Z^1 and Z^2 are i.i.d. Beta(a, b) random variables. To bound the expression in (4.11), we first condition on Z^1. By the tower rule, we have that

$$\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \leq \delta_{t_\star}\}} \mid T_{t_\star}] = \mathbb{E}[\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \leq \delta_{t_\star}\}} \mid Z^1, T_{t_\star}] \mid T_{t_\star}].$$

Conditioned on Z^1 and $T_{t_\star}$, we have that $Z^2 \sim \text{Beta}(a, b)$, so we can compute this conditional expectation explicitly:

$$\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \leq \delta_{t_\star}\}} \mid Z^1, T_{t_\star}] = \frac{1}{B(a, b)} \int_{(Z^1 - \delta_{t_\star})\vee 0}^{(Z^1 + \delta_{t_\star})\wedge 1} x^{a-1}(1 - x)^{b-1}\, dx, \tag{4.13}$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function. Recall from (4.10) that we only care about bounding this expression when the event $\mathcal{B}$ holds. From the definition of $\mathcal{B}$, and also the definitions of a and b (see (4.12)), it follows that if $\mathcal{B}$ holds, then $a, b > 2$ for all $t_\star$ large enough. We know that if $a, b > 1$, then the mode of the Beta(a,b) distribution is at $\frac{a-1}{a+b-2}$. Plugging this into (4.13), we obtain, for all $t_\star$ large enough, that

$$\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \le \delta_{t_\star}\}} \mid Z^1, T_{t_\star}]\mathbf{1}_{\mathcal{B}} \le \frac{2\delta_{t_\star}}{B(a,b)} \left(\frac{a-1}{a+b-2}\right)^{a-1} \left(\frac{b-1}{a+b-2}\right)^{b-1} \mathbf{1}_{\mathcal{B}}. \tag{4.14}$$

Now using the standard inequalities $\sqrt{2\pi} n^{n+1/2} e^n \le n! \le e n^{n+1/2} e^n$, which hold for all $n \ge 1$, we have that

$$B(a,b) = \frac{(a-1)!(b-1)!}{(a+b-1)!} \ge 2\pi \left(\frac{a-1}{a+b-1}\right)^{a-1} \left(\frac{b-1}{a+b-1}\right)^{b-1} \frac{\sqrt{(a-1)(b-1)}}{(a+b-1)^{3/2}}.$$

Therefore,

$$\begin{aligned} \frac{1}{B(a,b)} \left(\frac{a-1}{a+b-2}\right)^{a-1} \left(\frac{b-1}{a+b-2}\right)^{b-1} &\le \frac{1}{2\pi} \frac{(a+b-1)^{3/2}}{\sqrt{(a-1)(b-1)}} \left(1 + \frac{1}{a+b-2}\right)^{a+b-2} \\ &\le \frac{e}{2\pi} \frac{(a+b-1)^{3/2}}{\sqrt{(a-1)(b-1)}}. \end{aligned}$$

Plugging this back into (4.14), we obtain, for all $t_\star$ large enough, that

$$\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \le \delta_{t_\star}\}} \mid Z^1, T_{t_\star}]\mathbf{1}_{\mathcal{B}} \le C\delta_{t_\star} \frac{(a+b)^{3/2}}{\sqrt{ab}} \mathbf{1}_{\mathcal{B}}$$

for some constant C. From (4.12) we have that $a + b \le t_\star$. We also have that $b \ge t_\star/2 - 1/2$. Furthermore, on the event $\mathcal{B}$ we have that $a \ge t_\star/\sqrt{\log t_\star} - 1/2$. Altogether these imply that

$$\frac{(a+b)^{3/2}}{\sqrt{ab}} \mathbf{1}_{\mathcal{B}} \le C' t_\star^{1/2} \log^{1/4}(t_\star)$$

for some constant C' and all $t_\star$ large enough. Plugging this back into the previous display and using the definition of $\delta_{t_\star}$ we obtain that

$$\mathbb{E}[\mathbf{1}_{\{|Z^1 - Z^2| \le \delta_{t_\star}\}} \mid Z^1, T_{t_\star}]\mathbf{1}_{\mathcal{B}} \le \frac{C''}{\log^{1/4}(t_\star)}$$

for some constant C'' and all $t_\star$ large enough. Now taking an expectation over Z^1 and using (4.10) and (4.11), we finally obtain that

$$\limsup_{n\to\infty} \mathbb{P}(\{|X_n^1 - X_n^2| \le \delta_{t_\star}\} \cap \mathcal{A}^1 \cap \mathcal{A}^2 \cap \mathcal{B}) \le \frac{C''}{\log^{1/4}(t_\star)}$$

for all $t_\star$ large enough. This goes to zero as $t_\star \to \infty$, which concludes the proof. □

5. Estimating $t_\star$ with vanishing relative error as $t_\star \to \infty$. In this section we prove Theorem 1.9. To do this, we build on the ideas and the estimator introduced in Section 4, which provided an initial, coarse estimate of $t_\star$. The key additional idea compared to Section 4 is to average, over many subtrees, statistics similar to Y_n; see Figure 5 for an illustration. We start by defining precisely the estimator used to prove Theorem 1.9.

For a tree T_n on n vertices, let $T_n(k)$ denote the kth largest subtree of the rooted tree $(T_n, \theta(n))$ (with ties broken arbitrarily), with the root of this subtree denoted by $\widetilde{\theta}_n(k)$. In

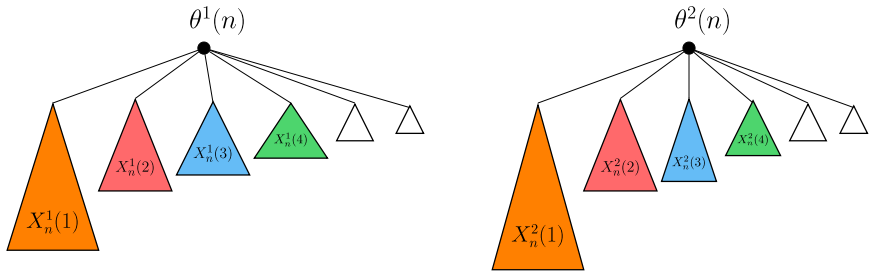


FIG. 5. *The pendent subtrees of the centroids $\theta^1(n)$ and $\theta^2(n)$ in T_n^1 and T_n^2, respectively, are ordered in decreasing order. The estimator studied in Section 5 matches several of the largest pendent subtrees in the two trees, as indicated by the colors in the figure.*

particular, with this notation we have that $\Psi_n(\theta(n)) = |T_n(1)|$. As before, for anything defined for a tree T_n, if we add a superscript i to it (where $i \in \{1, 2\}$), this means that it is the appropriate object in the tree T_n^i. For $i \in \{1, 2\}$ and $k \geq 1$, define the normalized subtree size

$$X_n^i(k) := \frac{1}{n}|T_n^i(k)|;$$

see Figure 5 for an illustration. Now define

$$Y_n(k) := \frac{(X_n^1(k) - X_n^2(k))^2}{2X_n^1(k)(1 - X_n^1(k))}$$

and note that $Y_n(1) \equiv Y_n$. For any $k \geq 1$ define

$$S_n(k) := \frac{1}{k}\sum_{\ell=1}^{k} Y_n(\ell).$$

For $k = 1$ we have that $S_n(1) = Y_n(1) = Y_n$ and everything proved in Section 4 applies. For $k > 1$ (and k not too large, to be made precise later), we still have that $S_n(k)$ is concentrated around $1/t_\star$. The improvement in $S_n(k)$ for large k, compared to $S_n(1)$, is that $S_n(k)$ has smaller variance than $S_n(1)$, by roughly a factor of order k.

In order to obtain a significant improvement over $S_n(1)$, we aim to use $S_n(k)$ with a choice of k that diverges as $t_\star \to \infty$. The catch is that $t_\star$ is unknown—in fact, it is the quantity that we desire to estimate. This is where it is useful to have an initial, coarse estimate of $t_\star$, which allows to choose an appropriate k. To this end, define

$$K_n := \left\lfloor -\frac{1}{400}\log Y_n \right\rfloor.$$

Our estimator for $t_\star$ is then

$$\widehat{t}_n := \frac{1}{S_n(K_n)}.$$

Theorem 1.9 then follows immediately from the following result.

THEOREM 5.1. *Let S be the tree on two vertices and let $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. We have that*

$$\lim_{t_\star\to\infty} \liminf_{n\to\infty} \mathbb{P}\left(\left(1 - \frac{\log\log t_\star}{2\sqrt{\log t_\star}}\right)\frac{1}{t_\star} \leq S_n(K_n) \leq \left(1 + \frac{\log\log t_\star}{\sqrt{\log t_\star}}\right)\frac{1}{t_\star}\right) = 1.$$

The same result also holds when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

In the remainder of this section, which is structured similar to Section 4, we prove this theorem. We start in Section 5.1 with some preliminaries: specifically, we define a couple of "nice" events on the space of sequences of growing trees, on which we will obtain bounds for $S_n(K_n)$. We state and prove first moment estimates in Section 5.2, where we also state a variance estimate whose proof we defer to Appendix B.2. We then prove Theorem 5.1 in Section 5.3, using the fact that the previously defined "nice" events have probability close to 1. We prove this latter fact in Appendix B.1.

5.1. *Preliminaries.* In Section 4 we defined "nice" events $\mathcal{A}$ and $\mathcal{B}$. Here, we define analogous "nice" events, which we denote by $\mathcal{C}$ and $\mathcal{D}$. First, we define

$$K \equiv K(t_\star) := \left\lfloor \frac{1}{384} \log t_\star \right\rfloor,$$

which we fix for the rest of Section 5. We are now ready to define the event $\mathcal{C}$.

DEFINITION 5.2 (The event $\mathcal{C}$). Given a sequence of trees $\{T_n\}_{n \geq t_\star}$, we say that the event $\mathcal{C}$ holds if and only if the following three properties all hold:

(C1) The centroid $\theta(n)$ is unique for all $n \geq t_\star$ and $\theta(n) = \theta(t_\star)$ for all $n \geq t_\star$.

(C2) For all integers $1 \leq k \leq K$, the vertex $\widetilde{\theta}_n(k)$ is uniquely defined for all $n \geq t_\star$, and also $\widetilde{\theta}_n(k) = \widetilde{\theta}_{t_\star}(k)$ for all $n \geq t_\star$.

(C3) For all $n \geq t_\star$ and all $1 \leq k \leq K$, we have that

$$\left| \frac{1}{n} \left| T_n(k) \right| - \frac{1}{t_\star} \left| T_{t_\star}(k) \right| \right| \leq \frac{1}{t_\star^{1/3}} \min\left\{ \frac{1}{t_\star} \left| T_{t_\star}(k) \right|, 1 - \frac{1}{t_\star} \left| T_{t_\star}(k) \right| \right\}. \tag{5.1}$$

As in Definition 4.3, the exponent $1/3$ in (5.1) is chosen for simplicity; any positive constant that is less than $1/2$ is a good choice for everything that follows (though the choice impacts the choice of other constants/exponents later on). Also, we always have that $|T_{t_\star}(k)| \leq |T_{t_\star}(1)| = \Psi_{t_\star}(\theta(t_\star)) \leq t_\star/2$, so the minimum in (5.1) is always attained by the first term; we include the second term in the definition just for clarity. Given a sequence of trees $\{T_n\}_{n \geq 2}$, we say that the event $\mathcal{C}$ holds if and only if it holds for the subsequence $\{T_n\}_{n \geq t_\star}$. The event $\mathcal{C}$ clearly depends on $t_\star$, but we choose to omit $t_\star$ from the notation in order to keep notation lighter. The following lemma shows that for PA and UA trees the event $\mathcal{C}$ holds with probability close to 1 when $t_\star$ is large.

LEMMA 5.3. *Let $\{T_n\}_{n \geq 2}$ be a sequence of trees started from the seed S and grown according to PA or UA. There exists a finite constant C such that for every $t_\star \geq 2$ we have that*

$$\mathbb{P}(\mathcal{C}^c) \leq \frac{C}{t_\star^{1/2000}}, \tag{5.2}$$

where $\mathcal{C}^c$ denotes the complement of $\mathcal{C}$.

Lemma 5.3 follows directly from Lemma 5.4 below.

Since the event $\mathcal{C}$ is analogous to the event $\mathcal{A}$, the intuition is similar. Let $\mathcal{C}^1$ and $\mathcal{C}^2$ denote the "nice" events corresponding to $\{T_n^1\}_{n \geq 2}$ and $\{T_n^2\}_{n \geq 2}$, respectively. The key point of the construction is that on the event $\mathcal{C}^1 \cap \mathcal{C}^2$, studying $X_n^1(k)$ and $X_n^2(k)$ reduces to studying the evolution of *fixed* subtrees that are present in the tree at time $t_\star$.

Formally, condition on the tree $T_{t_\star}^1 = T_{t_\star}^2 =: T_{t_\star}$. To abbreviate notation, we write $\theta := \theta(t_\star)$ and $\widetilde{\theta}(k) := \widetilde{\theta}_{t_\star}(k)$ for all $1 \le k \le K$; importantly, note that these are now *fixed* vertices (i.e., they do not change with n). Define the random variables

$$Z_n^i(k) := \frac{1}{n}\big|(T_n^i, \theta)_{\widetilde{\theta}(k)\downarrow}\big|$$

for $i \in \{1, 2\}$, $1 \le k \le K$, and $n \ge t_\star$. On the event $\mathcal{C}^i$ we have that $X_n^i(k) = Z_n^i(k)$ for all $n \ge t_\star$ and all $1 \le k \le K$.

As discussed in Section 4 for $k = 1$, by classical results on Pólya urns it follows that the limiting random variables

$$Z^i(k) := \lim_{n\to\infty} Z_n^i(k)$$

exist almost surely for $i \in \{1, 2\}$ and $1 \le k \le K$, for both PA and UA trees. Moreover, for any $1 \le k \le K$, we have that $Z^1(k)$ and $Z^2(k)$ are i.i.d. (this is conditional independence given $T_{t_\star}$) beta random variables, with parameters given as follows:

$$Z(k) \sim \begin{cases} \mathrm{Beta}\big(|T_{t_\star}(k)|, t_\star - |T_{t_\star}(k)|\big) & \text{for UA,} \\ \mathrm{Beta}\Big(|T_{t_\star}(k)| - \frac{1}{2}, t_\star - |T_{t_\star}(k)| - \frac{1}{2}\Big) & \text{for PA.} \end{cases} \tag{5.3}$$

Here $Z(k)$ is a random variable with the same distribution as $Z^1(k)$ and $Z^2(k)$.

From (5.3) it is clear that the quantity $|T_{t_\star}(k)|$ plays an important role in the distribution of $Z(k)$. In Section 4 we defined $\mathcal{B}$ to be the event that $|T_{t_\star}(1)| \ge t_\star/\sqrt{\log t_\star}$. Here we analogously want to define an event $\mathcal{D}$ on which we have lower bounds for $|T_{t_\star}(k)|$ for all $1 \le k \le K$. However, it turns out that we need some further properties from the event $\mathcal{D}$; because of this we do not define it explicitly here—see Appendix B.1 for an implicit definition. The following lemma guarantees the existence of an event $\mathcal{D}$ with the appropriate properties.

LEMMA 5.4. *Let $\{T_n\}_{n\ge 2}$ be a sequence of trees started from the seed S and grown according to PA or UA. There exists a finite constant C such that for every $t_\star \ge C$ the following holds. There exists a $T_{t_\star}$-measurable event $\mathcal{D}$ such that the following three things hold. First, on $\mathcal{D}$ we have for all $1 \le k \le K$ that*

$$|T_{t_\star}(k)| \ge t_\star^{7/8}.$$

Second,

$$\mathbb{P}(\mathcal{D}^c) \le \frac{C}{t_\star^{1/2000}}.$$

Finally,

$$\mathbb{P}(\mathcal{C}^c \mid \mathcal{D}) \le \frac{C}{t_\star^3}. \tag{5.4}$$

We note that the bound in (5.4) can be improved to a bound that decays faster than any polynomial in $t_\star$; however, we only state this simpler, weaker bound, since this is all we need for our purposes. The proof of Lemma 5.4 is deferred to Appendix B.1. In the following, $\mathcal{D}$ always refers to the event guaranteed by Lemma 5.4.

5.2. *First and second moment estimates.* We first state and prove the following first moment estimates.

LEMMA 5.5. *Let* $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. *For all* $t_\star$ *large enough the following holds. Fix* $k \in \{1, 2, \dots, K\}$. *Then, we have that*

$$\limsup_{n\to\infty} \mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, \mathcal{D}\big] \le \frac{1 + 3t_\star^{-1/3}}{t_\star} \tag{5.5}$$

and that

$$\liminf_{n\to\infty} \mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, \mathcal{D}\big] \ge \frac{1 - 3t_\star^{-1/3}}{t_\star}. \tag{5.6}$$

The same bounds also hold when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

PROOF. We start with the upper bound. By the exact same arguments as in the proof of Lemma 4.7, we have for every $\ell \in \{1, \dots, K\}$ that

$$\limsup_{n\to\infty} \mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, \mathcal{D}\big] \le \left(1 + \frac{1}{t_\star - 1}\right)^2 \big(1 - t_\star^{-1/3}\big)^{-2} \frac{1}{t_\star}.$$

Therefore, by linearity of expectation we also have that

$$\limsup_{n\to\infty} \mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, \mathcal{D}\big] \le \left(1 + \frac{1}{t_\star - 1}\right)^2 \big(1 - t_\star^{-1/3}\big)^{-2} \frac{1}{t_\star}.$$

The right hand side of the display above is at most $(1 + 3t_\star^{-1/3})/t_\star$ for all $t_\star$ large enough, which concludes the proof of (5.5).

We now turn to the lower bound. This follows similar lines as the upper bound, but an additional argument is needed. Fix $\ell \in \{1, \dots, K\}$. We again condition on the tree $T_{t_\star}$ at time $t_\star$; by the tower rule we have that

$$\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, \mathcal{D}\big] = \mathbb{E}\big[\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, T_{t_\star}\big] \,|\, \mathcal{D}\big].$$

Now given $T_{t_\star}$ such that $\mathcal{D}$ holds, property (C3) in Definition 5.2 implies that on the event $\mathcal{C}^1$ we have that

$$X_n^1(\ell)\big(1 - X_n^1(\ell)\big) \le \frac{1}{t_\star}\big|T_{t_\star}^1(\ell)\big|\left(1 - \frac{1}{t_\star}\big|T_{t_\star}^1(\ell)\big|\right)\big(1 + t_\star^{-1/3}\big)^2$$

for $n \ge t_\star$. Plugging this inequality into the definition of $Y_n(\ell)$ we obtain that

$$\begin{aligned}\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, T_{t_\star}\big] &\ge \frac{\mathbb{E}[(X_n^1(\ell) - X_n^2(\ell))^2\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, T_{t_\star}]}{2 \cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1 - \frac{1}{t_\star}|T_{t_\star}(\ell)|)(1 + t_\star^{-1/3})^2} \\ &= \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, T_{t_\star}]}{2 \cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1 - \frac{1}{t_\star}|T_{t_\star}(\ell)|)(1 + t_\star^{-1/3})^2},\end{aligned}$$

where the equality follows by observing that on the event $\mathcal{C}^1 \cap \mathcal{C}^2$ we have that $X_n^i(\ell) = Z_n^i(\ell)$ for $i \in \{1, 2\}$. Now writing the indicator as $\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} = 1 - \mathbf{1}_{(\mathcal{C}^1\cap\mathcal{C}^2)^c}$, we have that

$$\begin{aligned}\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,|\, T_{t_\star}\big] \ge{}& \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2 \,|\, T_{t_\star}]}{2 \cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1 - \frac{1}{t_\star}|T_{t_\star}(\ell)|)(1 + t_\star^{-1/3})^2} \\ &- \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2\mathbf{1}_{(\mathcal{C}^1\cap\mathcal{C}^2)^c} \,|\, T_{t_\star}]}{2 \cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1 - \frac{1}{t_\star}|T_{t_\star}(\ell)|)(1 + t_\star^{-1/3})^2}.\end{aligned} \tag{5.7}$$

We deal with the two terms in (5.7) separately, starting with the first term, for which the analysis is similar to that in the upper bound.

By the bounded convergence theorem we have that

$$\lim_{n\to\infty} \mathbb{E}\big[\big(Z_n^1(\ell) - Z_n^2(\ell)\big)^2 \,|\, T_{t_\star}\big] = \mathbb{E}\big[\big(Z^1(\ell) - Z^2(\ell)\big)^2 \,|\, T_{t_\star}\big].$$

Now using conditional independence, the limiting conditional distribution obtained in (5.3), and plugging in the variance of the beta distribution, we have that

$$\begin{aligned}&\mathbb{E}\big[\big(Z^1(\ell) - Z^2(\ell)\big)^2 \,|\, T_{t_\star}\big] \\ &\quad = 2\operatorname{Var}\big(Z(\ell) \,|\, T_{t_\star}\big) = \begin{cases} \dfrac{2|T_{t_\star}(\ell)|(t_\star - |T_{t_\star}(\ell)|)}{t_\star^2(t_\star+1)} & \text{for UA,} \\ \dfrac{2(|T_{t_\star}(\ell)| - 1/2)(t_\star - |T_{t_\star}(\ell)| - 1/2)}{(t_\star-1)^2 t_\star} & \text{for PA.} \end{cases}\end{aligned}$$

Plugging these formulas into the above, we obtain for UA trees that

$$\begin{aligned}\lim_{n\to\infty} \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2 \,|\, T_{t_\star}]}{2\cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1-\frac{1}{t_\star}|T_{t_\star}(\ell)|)(1+t_\star^{-1/3})^2} &= \frac{1}{(t_\star+1)(1+t_\star^{-1/3})^2} \\ &\geq \frac{1-2.5t_\star^{-1/3}}{t_\star},\end{aligned}$$

where the inequality holds for all $t_\star$ large enough. For PA trees we obtain that

$$\begin{aligned}&\lim_{n\to\infty} \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2 \,|\, T_{t_\star}]}{2\cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1-\frac{1}{t_\star}|T_{t_\star}(\ell)|)(1+t_\star^{-1/3})^2} \\ &\quad = \frac{1}{t_\star}\cdot\frac{1}{(1+t_\star^{-1/3})^2}\cdot\frac{t_\star^2}{(t_\star-1)^2}\cdot\frac{|T_{t_\star}(\ell)|-1/2}{|T_{t_\star}(\ell)|}\cdot\frac{t_\star - |T_{t_\star}(\ell)| - 1/2}{t_\star - |T_{t_\star}(\ell)|}.\end{aligned}$$

We always have that $|T_{t_\star}(\ell)| \leq t_\star/2$. Since $T_{t_\star}$ is such that $\mathcal{D}$ holds, by Lemma 5.4 we also have that $|T_{t_\star}(\ell)| \geq t_\star^{7/8}$. Plugging these inequalities into the display above, we obtain that

$$\begin{aligned}&\lim_{n\to\infty} \frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2 \,|\, T_{t_\star}]}{2\cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1-\frac{1}{t_\star}|T_{t_\star}(\ell)|)(1+t_\star^{-1/3})^2} \\ &\quad \geq \frac{1}{t_\star}\cdot\frac{1}{(1+t_\star^{-1/3})^2}\cdot\frac{t_\star^2}{(t_\star-1)^2}\cdot\frac{t_\star^{7/8}-1/2}{t_\star^{7/8}}\cdot\frac{t_\star/2-1/2}{t_\star/2} \geq \frac{1-2.5t_\star^{-1/3}}{t_\star},\end{aligned}$$

where the second inequality holds for all $t_\star$ large enough.

We now turn to the second term in (5.7). Since $Z_n^1(\ell) - Z_n^2(\ell) \in [-1,1]$, we have that

$$\mathbb{E}\big[\big(Z_n^1(\ell) - Z_n^2(\ell)\big)^2 \mathbf{1}_{(\mathcal{C}^1\cap\mathcal{C}^2)^c} \,|\, T_{t_\star}\big] \leq \mathbb{P}\big((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, T_{t_\star}\big).$$

As mentioned above, we always have that $|T_{t_\star}(\ell)| \leq t_\star/2$; moreover, since $T_{t_\star}$ is such that $\mathcal{D}$ holds, by Lemma 5.4 we also have that $|T_{t_\star}(\ell)| \geq t_\star^{7/8}$. Using these inequalities we may bound the second term in (5.7):

$$\frac{\mathbb{E}[(Z_n^1(\ell) - Z_n^2(\ell))^2 \mathbf{1}_{(\mathcal{C}^1\cap\mathcal{C}^2)^c} \,|\, T_{t_\star}]}{2\cdot \frac{1}{t_\star}|T_{t_\star}(\ell)|(1-\frac{1}{t_\star}|T_{t_\star}(\ell)|)(1+t_\star^{-1/3})^2} \leq t_\star^{1/8}\mathbb{P}\big((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, T_{t_\star}\big).$$

Taking an expectation over $T_{t_\star}$, this bound becomes $t_\star^{1/8}\mathbb{P}((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, \mathcal{D})$. By Lemma 5.4 we have that $\mathbb{P}((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, \mathcal{D}) \leq Ct_\star^{-3}$ for some finite constant C and all $t_\star$ large enough. Thus ultimately the bound becomes $Ct_\star^{-23/8}$, which is at most $0.5t_\star^{-4/3}$ for all $t_\star$ large enough.

Overall, we have thus shown that

$$\liminf_{n\to\infty} \mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \mid \mathcal{D}\big] \geq \frac{1-3t_\star^{-1/3}}{t_\star}$$

for all $t_\star$ large enough (where here "large enough" does not depend on ℓ). The bound in (5.6) follows by linearity of expectation. □

The following lemma gives a variance bound that we will use.

LEMMA 5.6. *Let $(T_n^1, T_n^2) \sim \mathrm{CPA}(n, t_\star, S)$. There exists a finite constant C such that for all $t_\star$ large enough we have for all $k \in \{1, 2, \ldots, K\}$ that*

$$\limsup_{n\to\infty} \mathrm{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \mid \mathcal{D}\big) \leq \frac{C}{kt_\star^2}. \tag{5.8}$$

The same bound also holds when $(T_n^1, T_n^2) \sim \mathrm{CUA}(n, t_\star, S)$.

The proof of Lemma 5.6 is somewhat lengthy, so we defer it to Appendix B.2.

5.3. *Putting everything together*: *Proof of Theorem 5.1.* PROOF OF THEOREM 5.1. In the following we set

$$\varepsilon := \frac{\log\log t_\star}{2\sqrt{\log t_\star}} \tag{5.9}$$

to abbreviate notation. Our goal is to show that

$$\lim_{t_\star\to\infty} \limsup_{n\to\infty} \mathbb{P}\left(\left|S_n(K_n) - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star}\right) = 0. \tag{5.10}$$

To do this, we first fix $k \in \{1, 2, \ldots, K\}$ and bound the probability $\mathbb{P}(|S_n(k) - \frac{1}{t_\star}| \geq \frac{\varepsilon}{t_\star})$. By conditioning on the "nice" event $\mathcal{D}$, we have that

$$\mathbb{P}\left(\left|S_n(k) - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star}\right) \leq \mathbb{P}\left(\left|S_n(k) - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star} \,\middle|\, \mathcal{D}\right) + \mathbb{P}(\mathcal{D}^c).$$

The second term above is at most $C/t_\star^{1/2000}$ by Lemma 5.4. We can break the first term above into two further terms, based on whether the "nice" event $\mathcal{C}^1 \cap \mathcal{C}^2$ holds or not: by a union bound we have that

$$\mathbb{P}\left(\left|S_n(k) - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star} \,\middle|\, \mathcal{D}\right) \leq \mathbb{P}\left(\left|S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star} \,\middle|\, \mathcal{D}\right) + \mathbb{P}\big((\mathcal{C}^1 \cap \mathcal{C}^2)^c \mid \mathcal{D}\big).$$

The second term in the display above is at most $C/t_\star^3$ by Lemma 5.4, so it remains to deal with the first term above. Recall that Lemma 5.5 implies that for all $t_\star$ large enough we have for all n large enough that

$$\left|\mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \mid \mathcal{D}\big] - \frac{1}{t_\star}\right| \leq \frac{4t_\star^{-1/3}}{t_\star}.$$

Recalling the definition of ε from (5.9), note that $\varepsilon \geq 8t_\star^{-1/3}$ for all $t_\star$ large enough and hence $\varepsilon - 4t_\star^{-1/3} \geq \varepsilon/2$ for all $t_\star$ large enough. By the triangle inequality we thus have that

$$\begin{aligned}&\mathbb{P}\left(\left|S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} - \frac{1}{t_\star}\right| \geq \frac{\varepsilon}{t_\star} \,\middle|\, \mathcal{D}\right)\\ &\quad\leq \mathbb{P}\left(\left|S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} - \mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \mid \mathcal{D}\big]\right| \geq \frac{\varepsilon}{2t_\star} \,\middle|\, \mathcal{D}\right).\end{aligned}$$

Finally, by Chebyshev's inequality we have that

$$\mathbb{P}\left(\left|S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} - \mathbb{E}[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,|\,\mathcal{D}]\right| \geq \frac{\varepsilon}{2t_\star}\,\middle|\,\mathcal{D}\right) \leq \frac{4t_\star^2}{\varepsilon^2}\operatorname{Var}(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,|\,\mathcal{D}).$$

Taking a limit as $n\to\infty$ and putting all the above bounds together we have thus obtained that

$$\limsup_{n\to\infty}\mathbb{P}\left(\left|S_n(k)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right)\leq\frac{C}{k\varepsilon^2}+\frac{C}{t_\star^{1/2000}} \tag{5.11}$$

for some finite constant C and all $t_\star$ large enough.

Now we are ready to show (5.10). Define the event

$$\mathcal{E} := \{\log t_\star - \log\log t_\star \leq -\log Y_n \leq \log t_\star + \log\log t_\star\}.$$

By a union bound we have that

$$\mathbb{P}\left(\left|S_n(K_n)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right)\leq\mathbb{P}\left(\left\{\left|S_n(K_n)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right\}\cap\mathcal{E}\right)+\mathbb{P}(\mathcal{E}^c).$$

By Theorem 4.2 we have that $\lim_{t_\star\to\infty}\limsup_{n\to\infty}\mathbb{P}(\mathcal{E}^c)=0$, so what remains is to deal with the first term on the right hand side of the display above. On the event $\mathcal{E}$ we have that

$$\left\lfloor\frac{1}{400}\log t_\star-\frac{1}{400}\log\log t_\star\right\rfloor\leq K_n\leq\left\lfloor\frac{1}{400}\log t_\star+\frac{1}{400}\log\log t_\star\right\rfloor,$$

so by a union bound we have that

$$\mathbb{P}\left(\left\{\left|S_n(K_n)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right\}\cap\mathcal{E}\right)\leq\sum_{k=\lfloor\frac{1}{400}\log t_\star-\frac{1}{400}\log\log t_\star\rfloor}^{\lfloor\frac{1}{400}\log t_\star+\frac{1}{400}\log\log t_\star\rfloor}\mathbb{P}\left(\left|S_n(k)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right).$$

Note that $\frac{1}{400}\log t_\star-\frac{1}{400}\log\log t_\star\leq K$ for all $t_\star$ large enough, so we can apply the bound (5.11) that holds for fixed $k\leq K$. Thus taking a limit as $n\to\infty$ and applying (5.11) we thus obtain that

$$\limsup_{n\to\infty}\mathbb{P}\left(\left\{\left|S_n(K_n)-\frac{1}{t_\star}\right|\geq\frac{\varepsilon}{t_\star}\right\}\cap\mathcal{E}\right)\leq\frac{C\log\log t_\star}{\varepsilon^2\log t_\star}+\frac{C\log\log t_\star}{t_\star^{1/2000}}\leq\frac{C'}{\log\log t_\star}$$

for some finite constants C and C', and all $t_\star$ large enough, where in the second inequality we used the definition of ε from (5.9). Taking the limit as $t_\star\to\infty$ concludes the proof. □

APPENDIX A: PROOFS OF REMAINING LEMMAS IN SECTION 4

In this subsection we prove Lemmas 4.4 and 4.6, proofs that we have deferred until now.

A.1. Proof of Lemma 4.6. We start with the proof of Lemma 4.6, which is relatively short.

PROOF OF LEMMA 4.6. First, by a union bound we have that

$$\mathbb{P}(\mathcal{B}^c)=\mathbb{P}\left(\Psi_{t_\star}(\theta(t_\star))<\frac{t_\star}{\sqrt{\log t_\star}}\right)\leq\sum_{i=1}^{t_\star}\mathbb{P}\left(\Psi_{t_\star}(v_i)<\frac{t_\star}{\sqrt{\log t_\star}}\right). \tag{A.1}$$

Noting that the term for $i=1$ is equal to the term for $i=2$, we now fix $i\geq 2$. Note that v_i is introduced in T_i. Let w denote the neighbor of v_i in T_i. By definition we have that

$$\Psi_{t_\star}(v_i)=\max_{u\in\mathcal{N}_{v_i}(T_{t_\star})}\left|(T_{t_\star},v_i)_{u\downarrow}\right|\geq\left|(T_{t_\star},v_i)_{w\downarrow}\right|$$

and so—introducing $M_n := \frac{1}{n}|(T_n, v_i)_{w\downarrow}|$ for $n \geq i$ in order to abbreviate notation—we have that

$$\mathbb{P}\left(\Psi_{t_\star}(v_i) < \frac{t_\star}{\sqrt{\log t_\star}}\right) \leq \mathbb{P}\left(M_{t_\star} \leq \frac{1}{\sqrt{\log t_\star}}\right).$$

This latter probability can be understood using Pólya urn and martingale arguments. The proofs for PA and UA trees are similar, and we start with UA trees. For UA trees, the evolution of the pair $(nM_n, n - nM_n)$ for $n \geq i$ follows a classical Pólya urn with initial condition $(i-1, 1)$. By standard results on Pólya urns we have that $\{M_n\}_{n\geq i}$ is a martingale, the limit $M_\infty := \lim_{n\to\infty} M_n$ exists almost surely, and $M_\infty \sim \text{Beta}(i-1, 1)$. By this latter property we have that

$$\mathbb{P}(M_\infty \leq z) = z^{i-1} \tag{A.2}$$

for all $z \in (0, 1)$. Since $\{M_n\}_{n\geq i}$ is a nonnegative martingale, we also have that

$$\mathbb{P}(M_\infty \leq 2z \,|\, M_n \leq z) \geq 1/2$$

for all $z \geq 0$ and $n \geq i$, which implies that $\mathbb{P}(M_n \leq z) \leq 2\mathbb{P}(M_\infty \leq 2z)$. Thus using (A.2) we have that

$$\mathbb{P}\left(M_{t_\star} \leq \frac{1}{\sqrt{\log t_\star}}\right) \leq 2\left(\frac{2}{\sqrt{\log t_\star}}\right)^{i-1}$$

for all $t_\star$ large enough. Plugging this bound back into (A.1) and noting that the geometric sum is on the same order as the largest term, we obtain that

$$\mathbb{P}(\mathcal{B}^c) \leq \frac{12}{\sqrt{\log t_\star}}$$

for all $t_\star$ large enough.

Turning now to PA trees, the evolution of the pair $(2nM_n - 1, 2n - 2nM_n - 1)$ for $n \geq i$ follows a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$ and initial condition $(2i-3, 1)$. Define $\widetilde{M}_n := (2nM_n - 1)/(2n-2)$. The process $\{\widetilde{M}_n\}_{n\geq i}$ is a bounded martingale and hence its limit as $n \to \infty$ exists almost surely. Since

$$\widetilde{M}_n = M_n + \frac{1}{n-1}M_n - \frac{1}{2n-2}$$

and $M_n \in [0, 1]$, the limit of the martingale equals the limit of M_n; that is, $M_\infty := \lim_{n\to\infty} M_n = \lim_{n\to\infty} \widetilde{M}_n$ exists almost surely. Furthermore, by standard results on Pólya urns we know that $M_\infty \sim \text{Beta}(i - 3/2, 1/2)$. By this latter property, and using the bound $(1-z)^{-1/2} \leq \sqrt{2}$ for $z \in (0, 1/2)$ in the density function of the beta distribution, we have that

$$\mathbb{P}(M_\infty \leq z) \leq \frac{\sqrt{2}}{(i-\frac{3}{2})B(i-\frac{3}{2}, \frac{1}{2})} z^{i-3/2}$$

for all $z \in (0, 1/2)$. We can further bound this quantity using properties of the Gamma function. Specifically, we use the following identities: $\Gamma(z+1) = z\Gamma(z)$, for a positive integer n we have that $\Gamma(n+1) = n!$ and also that $\Gamma(n+1/2) = \frac{(2n)!\sqrt{\pi}}{4^n n!}$, and finally that $\Gamma(1/2) = \sqrt{\pi}$. Using these we have that

$$\left(i - \frac{3}{2}\right)B\left(i - \frac{3}{2}, \frac{1}{2}\right) = \frac{(i-\frac{3}{2})\Gamma(i-\frac{3}{2})\Gamma(\frac{1}{2})}{\Gamma(i-1)} = \pi(i-1)\binom{2i-2}{i-1}4^{-i+1} \geq 4^{-i+1}. \tag{A.3}$$

Plugging this back into the previous display we obtain that

$$\mathbb{P}(M_\infty \le z) \le 4(4z)^{i-3/2}$$

for all $z \in (0, 1/2)$. Using the fact that $\widetilde{M}_n \le 2M_n$, together with the same martingale argument as before, we have that

$$\mathbb{P}(M_n \le z) \le \mathbb{P}(\widetilde{M}_n \le 2z) \le 2\mathbb{P}(M_\infty \le 4z).$$

The previous two displays combined imply that $\mathbb{P}(M_n \le z) \le 8(16z)^{i-3/2}$ for all $z \in (0, 1/8)$ and $n \ge i$. We have thus obtained that

$$\mathbb{P}\biggl(M_{t_\star} \le \frac{1}{\sqrt{\log t_\star}}\biggr) \le 8\biggl(\frac{16}{\sqrt{\log t_\star}}\biggr)^{i-3/2}$$

for all $t_\star$ large enough. Plugging this bound back into (A.1) and noting that the geometric sum is on the same order as the largest term, we obtain the desired bound (4.6). □

A.2. Proof of Lemma 4.4. We now turn to the proof of Lemma 4.4, which is more involved. We start by stating and proving a few auxiliary lemmas that we will use.

The following lemma gives us an exponential bound on the probability that a vertex of large timestamp ever becomes the centroid. This was proved in [27]; see their Lemmas A.1 and 3.1.

LEMMA A.1. *Consider a sequence of PA or UA trees started from the seed $S = S_2$. For all t large enough we have that*

$$\mathbb{P}\bigl(v_{t+1} \textit{ becomes at least as central as } \theta(t) \textit{ at some future time}\bigr) \le \frac{P(t/2)}{2^{t/2}},$$

where P is a fixed polynomial.

The following lemma is useful in studying the relative (anti-)centralities of two vertices by examining the growth of specific subtrees.

LEMMA A.2. *Let $\{T_n\}_{n\ge 2}$ be a sequence of growing trees (such as PA or UA trees), where at every time step a single vertex is added to the tree, together with a single edge. Let $v_1, v_2, v_3, \ldots$ denote the vertices in order of appearance. Fix t and let i and j be distinct positive integers such that $i, j \le t$. Suppose that*

$$\Psi_t(v_i) > \Psi_t(v_j) \tag{A.4}$$

and that there exists $N > t$ such that

$$\Psi_N(v_i) \le \Psi_N(v_j). \tag{A.5}$$

Then there must exist M such that $t < M \le N$ and

$$\bigl|(T_M, v_i)_{v_j\downarrow}\bigr| = \bigl|(T_M, v_j)_{v_i\downarrow}\bigr|.$$

PROOF. We start with some notation. Fix $n \ge t$ and consider the tree T_n. Let $a_1, a_2, a_3, \ldots$ denote the sizes of the pendent subtrees of v_i, excluding the subtree that contains v_j. Similarly, let $b_1, b_2, b_3, \ldots$ denote the sizes of the pendent subtrees of v_j, excluding the subtree that contains v_i. Finally, let c denote the number of vertices that are "in between" v_i and v_j; that is, c is the number of vertices u such that the path from u to v_i does not contain v_j and

the path from u to v_j does not contain v_i. Note that with this notation we have the following equalities:

$$\Psi_n(v_i) = \max\Big\{c + 1 + \sum_\ell b_\ell, a_1, a_2, a_3, \dots\Big\},$$

$$\Psi_n(v_j) = \max\Big\{c + 1 + \sum_k a_k, b_1, b_2, b_3, \dots\Big\},$$

and also

$$|(T_n, v_i)_{v_j\downarrow}| = 1 + \sum_\ell b_\ell, \qquad |(T_n, v_j)_{v_i\downarrow}| = 1 + \sum_k a_k. \tag{A.6}$$

We now claim that if $\Psi_n(v_i) > \Psi_n(v_j)$, then $\sum_\ell b_\ell > \sum_k a_k$. We prove this by contradiction; suppose that $\sum_\ell b_\ell \le \sum_k a_k$. Then $\Psi_n(v_j) \ge c + 1 + \sum_k a_k \ge c + 1 + \sum_\ell b_\ell$, so $\Psi_n(v_i) > \Psi_n(v_j)$ implies that $\Psi_n(v_i) = a_{k'}$ for some k'. But then $\Psi_n(v_j) \ge c + 1 + \sum_k a_k \ge 1 + a_{k'} > \Psi_n(v_i)$, which is a contradiction. The same argument shows that if $\Psi_n(v_i) \ge \Psi_n(v_j)$, then $\sum_\ell b_\ell \ge \sum_k a_k$. As a corollary, we have that if $\Psi_n(v_i) = \Psi_n(v_j)$, then $\sum_\ell b_\ell = \sum_k a_k$.

Altogether, using (A.6), we have shown that

$$\operatorname{sgn}\big(\Psi_n(v_i) - \Psi_n(v_j)\big) = \operatorname{sgn}\big(|(T_n, v_i)_{v_j\downarrow}| - |(T_n, v_j)_{v_i\downarrow}|\big), \tag{A.7}$$

where sgn is the sign function: $\operatorname{sgn}(x) = -1$ if $x < 0$, $\operatorname{sgn}(x) = 0$ if $x = 0$, and $\operatorname{sgn}(x) = 1$ if $x > 0$. Observe also that the integer-valued quantity

$$f(n) := |(T_n, v_i)_{v_j\downarrow}| - |(T_n, v_j)_{v_i\downarrow}|$$

changes by 1, 0, or -1 as n increases by one. The assumption (A.4), together with (A.7), implies that $f(t) > 0$. The assumption (A.5), together with (A.7), implies that $f(N) \le 0$. Therefore, by the previous observation, there must exist $M \in \{t + 1, \dots, N\}$ such that $f(M) = 0$. □

The following lemma gives concentration bounds for Pólya urns.

LEMMA A.3. *Let $\{(A_n, B_n)\}_{n\ge0}$ be a stochastic process with a deterministic initial condition satisfying $A_0, B_0 \ge 1$, and let $k := A_0 + B_0$.*

If $\{(A_n, B_n)\}_{n\ge0}$ evolves as a classical Pólya urn, then for any $\varepsilon > 0$ we have that

$$\mathbb{P}\Big(\exists n \ge 0 : \Big|\frac{A_n}{k+n} - \frac{A_0}{k}\Big| \ge \varepsilon\Big) \le 2\exp(-k\varepsilon^2/2).$$

If $\{(2A_n - 1, 2B_n - 1)\}_{n\ge0}$ evolves as a Pólya urn with replacement matrix $\left(\begin{smallmatrix}2&0\\0&2\end{smallmatrix}\right)$, then for any $\varepsilon \ge 2/(k-1)$ we have that

$$\mathbb{P}\Big(\exists n \ge 0 : \Big|\frac{A_n}{k+n} - \frac{A_0}{k}\Big| \ge \varepsilon\Big) \le 2\exp(-(k-1)\varepsilon^2/8).$$

PROOF. We start with the first claim. Defining $M_n := A_n/(k+n)$, we have that $\{M_n\}_{n\ge0}$ is a martingale. The martingale differences satisfy $|M_n - M_{n-1}| \le 1/(k+n)$ for every $n \ge 1$. Therefore, by the maximal version of Azuma's inequality we have for every $\varepsilon > 0$ that

$$\mathbb{P}\Big(\sup_{n\ge0} |M_n - M_0| \ge \varepsilon\Big) \le 2\exp\Big(-\frac{\varepsilon^2}{2\sum_{n\ge1}(k+n)^{-2}}\Big).$$

The claim follows from the fact that $\sum_{n\ge1}(k+n)^{-2} \le 1/k$.

Turning to the second claim, first note that again $A_n + B_n = k + n$ for every $n \geq 0$. Define $\widetilde{M}_n := (2A_n - 1)/(2A_n + 2B_n - 2) = (2A_n - 1)/(2(k+n-1))$ and observe that the process $\{\widetilde{M}_n\}_{n\geq 0}$ is a martingale. Furthermore, the martingale differences satisfy $|\widetilde{M}_n - \widetilde{M}_{n-1}| \leq 1/(k+n-1)$ for every $n \geq 1$. Therefore, by the same argument as above we have for every $\varepsilon > 0$ that

$$\mathbb{P}\Big(\sup_{n\geq 0} |\widetilde{M}_n - \widetilde{M}_0| \geq \varepsilon\Big) \leq 2\exp\big(-(k-1)\varepsilon^2/2\big).$$

Now observe that $|(M_n - M_0) - (\widetilde{M}_n - \widetilde{M}_0)| \leq 1/(k-1)$, so by the triangle inequality we have that

$$\begin{aligned}\mathbb{P}\Big(\sup_{n\geq 0} |M_n - M_0| \geq \varepsilon\Big) &\leq \mathbb{P}\Big(\sup_{n\geq 0} |\widetilde{M}_n - \widetilde{M}_0| \geq \varepsilon - 1/(k-1)\Big)\\ &\leq \mathbb{P}\Big(\sup_{n\geq 0} |\widetilde{M}_n - \widetilde{M}_0| \geq \varepsilon/2\Big)\end{aligned}$$

for any $\varepsilon \geq 2/(k-1)$. The result follows by putting the previous two displays together. □

Finally, the following lemma gives a tail bound for degrees in PA and UA trees.

LEMMA A.4. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed $S = S_2$ and grown according to PA or UA. Let $v_1, v_2, v_3, \ldots$ denote the vertices in order of appearance. Let $d_n(v)$ denote the degree of v in T_n. There exists a positive constant c such that for every $1 \leq i \leq n$ we have that*

$$\mathbb{P}\big(d_n(v_i) \geq \sqrt{n}\log^2(n)\big) \leq \exp\big(-c\log^3(n)\big).$$

PROOF. The vertex v_3 attaches to either v_1 or v_2; without loss of generality, assume that it attaches to v_1, that is, v_1 has degree 2 in T_3. For both PA and UA trees, $d_n(v_1)$ stochastically dominates $d_n(v_i)$ for $1 < i \leq n$, so it suffices to prove the claim for v_1. Furthermore, the random variable $d_n(v_1)$ in a PA tree stochastically dominates the random variable $d_n(v_1)$ in a UA tree, hence it suffices to prove the claim for PA trees.

For $n \geq 3$ let $M_n := d_n(v_1)/\sqrt{n-2}$. Observe that

$$\mathbb{E}\big[d_{n+1}(v_1) \,|\, d_n(v_1)\big] = \Big(1 + \frac{1}{2n-2}\Big) d_n(v_1).$$

Since $(1 + 1/(2n-2))/\sqrt{n-1} \leq 1/\sqrt{n-2}$ for every $n \geq 3$, it follows that $\{M_n\}_{n\geq 3}$ is a supermartingale. Also, $|M_n - M_{n-1}| \leq 1/\sqrt{n-1}$. Thus by Azuma's inequality for supermartingales, noting that $M_3 = 2$, we have for every $\lambda > 0$ that

$$\mathbb{P}\Big(\frac{d_n(v_1)}{\sqrt{n-2}} - 2 \geq \lambda\Big) \leq \exp\Big(-\frac{\lambda^2}{2\sum_{i=4}^{n} 1/(i-1)}\Big) \leq \exp\Big(-\frac{\lambda^2}{2\log n}\Big).$$

Plugging in $\lambda = \log^2(n)$ yields the desired claim. □

We are now ready to prove Lemma 4.4.

PROOF OF LEMMA 4.4. We divide the proof into six steps. In the following we informally call a vertex an "early" vertex if its timestamp is at most $\log t_\star$.

Step 1: *The centroid is an early vertex.*

For a fixed $i \geq 1$, let $\mathcal{A}_1(i)$ denote the event that v_i never becomes a centroid during the whole process; that is, the event that v_i is not a centroid in T_s for any $s \geq i$. Define

$$\mathcal{A}_1 := \bigcap_{i > \log t_\star} \mathcal{A}_1(i).$$

An immediate consequence of Lemma A.1 is that $\mathbb{P}(\mathcal{A}_1(i)^c) \le \exp(-i/3)$ for all i large enough. So by a union bound we have, for all $t_\star$ large enough that

$$\mathbb{P}(\mathcal{A}_1^c) \le \sum_{i > \log t_\star} \mathbb{P}(\mathcal{A}_1(i)^c) \le \sum_{i > \log t_\star} e^{-i/3} \le \frac{4}{t_\star^{1/3}}.$$

Step 2: *Early subtrees are large in $T_{t_\star}$.*

This is an important intermediate step towards the overarching goal of characterizing the centroid. Specifically, the consequence of early subtree sizes being large is that then many of the random variables we will consider in future steps will be "stable" in timesteps $t \ge t_\star$.

For distinct positive integers $i, j \le \log t_\star$, we will show that subtrees of the form $(T_{t_\star}, v_i)_{v_j\downarrow}$ are large. Formally, for distinct positive integers $i, j \le \log t_\star$, define the event

$$\mathcal{E}_2(i, j) := \left\{ |(T_{t_\star}, v_i)_{v_j\downarrow}| \ge \frac{t_\star}{\log^7(t_\star)} \right\}.$$

We proceed by bounding the probability of the complement of $\mathcal{E}_2(i, j)$, using arguments similar to those found in the proof of Lemma 4.6. Since the details are repetitive, we only give the final bounds and leave the details to the reader.

Assume in the following that $1 \le i < j \le \log t_\star$. We start with UA trees. Let $\varphi_j \sim \text{Beta}(1, j-1)$. Then, by combining Pólya urn and martingale arguments as in the proof of Lemma 4.6, we have for every $z \in [0, 1]$ that

$$\max\left\{ \mathbb{P}\left(\frac{1}{t_\star} |(T_{t_\star}, v_i)_{v_j\downarrow}| \le z \right), \mathbb{P}\left(\frac{1}{t_\star} |(T_{t_\star}, v_j)_{v_i\downarrow}| \le z \right) \right\} \le 2\mathbb{P}(\varphi_j \le 2z).$$

For every $z \in [0, 1]$ we have that

$$\mathbb{P}(\varphi_j \le z) = (j-1) \int_0^z (1-x)^{j-2}\, dx \le (j-1)z.$$

Combining the previous two displays and using the fact that $j \le \log t_\star$, we have that

$$\max\{\mathbb{P}(\mathcal{E}_2(i, j)^c), \mathbb{P}(\mathcal{E}_2(j, i)^c)\} \le \frac{4j}{\log^7(t_\star)} \le \frac{4}{\log^6(t_\star)}.$$

Turning now to PA trees, let $\varphi'_j \sim \text{Beta}(1/2, j-3/2)$. Then, again by combining Pólya urn and martingale arguments as in the proof of Lemma 4.6, we have for every $z \in [0, 1]$ that

$$\max\left\{ \mathbb{P}\left(\frac{1}{t_\star} |(T_{t_\star}, v_i)_{v_j\downarrow}| \le z \right), \mathbb{P}\left(\frac{1}{t_\star} |(T_{t_\star}, v_j)_{v_i\downarrow}| \le z \right) \right\} \le 2\mathbb{P}(\varphi'_j \le 4z). \tag{A.8}$$

We have that

$$\mathbb{P}(\varphi'_j \le z) = \frac{1}{B(\frac{1}{2}, j-\frac{3}{2})} \int_0^z x^{-1/2}(1-x)^{j-5/2}\, dx \le \frac{2\sqrt{2}}{B(\frac{1}{2}, j-\frac{3}{2})} \sqrt{z},$$

where the inequality holds for every $z \in (0, 1/2)$. From (A.3) and the symmetry of the beta function we have that

$$B\left(\frac{1}{2}, j-\frac{3}{2}\right) = \pi \frac{j-1}{j-\frac{3}{2}} \binom{2j-2}{j-1} 4^{-j+1} \ge \frac{1}{\sqrt{j-1}}, \tag{A.9}$$

where the inequality follows by using the bound $\binom{2n}{n} \ge 4^n/\sqrt{4n}$ which holds for all $n \ge 1$. Combining the two previous displays we have obtained that $\mathbb{P}(\varphi'_j \le z) \le 2\sqrt{2jz}$ for all $z \in (0, 1/2)$. Plugging this back into (A.8) and using the fact that $j \le \log t_\star$, we have, for all $t_\star$ large enough, that

$$\max\{\mathbb{P}(\mathcal{E}_2(i, j)^c), \mathbb{P}(\mathcal{E}_2(j, i)^c)\} \le \frac{8\sqrt{2}\sqrt{j}}{\log^{7/2}(t_\star)} \le \frac{12}{\log^3(t_\star)}.$$

Altogether we have shown in this step that for all distinct positive integers $i, j \le \log t_\star$, and for both PA and UA trees, we have, for all $t_\star$ large enough, that

$$\mathbb{P}(\mathcal{E}_2(i,j)^c) \le \frac{12}{\log^3(t_\star)}.$$

Step 3: *The anti-centrality rankings for the early vertices are stable.*

Using Step 2, we will now show that the relative anti-centrality of any pair of early vertices is "stable" (with probability close to 1); that is, it does not change after a certain time. More specifically, we will show, for distinct positive integers $i, j \le \log t_\star$, that if $\Psi_{t_\star}(v_i) > \Psi_{t_\star}(v_j)$, then $\Psi_t(v_i) > \Psi_t(v_j)$ for every $t \ge t_\star$, with probability close to 1 (and similarly if the inequality goes the other way). We thus define the events

$$\mathcal{A}_3(i,j) := \{\forall t \ge t_\star : (\Psi_{t_\star}(v_i) - \Psi_{t_\star}(v_j))(\Psi_t(v_i) - \Psi_t(v_j)) > 0\}$$

for distinct positive integers $i, j \le \log t_\star$, and also

$$\mathcal{A}_3 := \bigcap_{\substack{1 \le i,j \le \log t_\star \\ i \ne j}} \mathcal{A}_3(i,j).$$

By Lemma A.2, if we wish to compare $\Psi_t(v_i)$ and $\Psi_t(v_j)$, it suffices to compare the sizes of the subtrees $(T_t, v_i)_{v_j\downarrow}$ and $(T_t, v_j)_{v_i\downarrow}$. This motivates defining the event

$$\mathcal{E}_3(i,j) := \left\{\forall t \ge t_\star : \left| \frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} - \frac{1}{2} \right| > \frac{1}{\log^3(t_\star)} \right\} \cap \mathcal{E}_2(i,j) \cap \mathcal{E}_2(j,i)$$

for distinct positive integers $i, j \le \log t_\star$. We claim that, for all $t_\star$ large enough, if $\mathcal{E}_3(i,j)$ holds, then $\mathcal{A}_3(i,j)$ must also hold. To see this, first note that on $\mathcal{E}_2(i,j) \cap \mathcal{E}_2(j,i)$ we have that

$$|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}| \ge |(T_{t_\star}, v_i)_{v_j\downarrow}| + |(T_{t_\star}, v_j)_{v_i\downarrow}| \ge \frac{t_\star}{\log^7(t_\star)}. \tag{A.10}$$

Since the quantity $|(T_t, v_i)_{v_j\downarrow}|$ can change by at most 1 at a time, the display above implies that the ratio

$$\frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} \tag{A.11}$$

can only change by at most $\log^7(t_\star)/t_\star$ at each time step. Since this is smaller than $1/\log^3(t_\star)$ for all $t_\star$ large enough, the event $\mathcal{E}_3(i,j)$ thus implies, for all $t_\star$ large enough, that the ratio in (A.11) is either strictly greater than $1/2$ for all $t \ge t_\star$ or strictly smaller than $1/2$ for all $t \ge t_\star$. In light of Lemma A.2, this implies that $\mathcal{A}_3(i,j)$ holds for all $t_\star$ large enough.

In the remainder of this step we thus focus on bounding the probability of $\mathcal{E}_3(i,j)$. Since $\mathcal{E}_3(i,j) = \mathcal{E}_3(j,i)$, we may, and thus will, assume in the following that $1 \le i < j \le \log t_\star$. To abbreviate notation, we introduce $J := |(T_j, v_j)_{v_i\downarrow}|$, and note that $1 \le J \le j-1$. We first give the proof for UA trees and subsequently explain what changes for PA trees.

Conditioned on T_j, the pair

$$\big(|(T_t, v_i)_{v_j\downarrow}|, |(T_t, v_j)_{v_i\downarrow}|\big),$$

when viewed at times when one of the coordinates increases, evolves as a classical Pólya urn started from $(1, J)$. Therefore, conditioned on T_j, the limit

$$\varphi_{i,j} := \lim_{t\to\infty} \frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} \tag{A.12}$$

exists almost surely, and moreover $\varphi_{i,j} \sim \mathrm{Beta}(1, J)$. Since this holds for every tree T_j on j vertices, the limiting random variable $\varphi_{i,j}$ exists almost surely unconditionally (and its distribution is a mixture of beta distributions). Plugging in the density of the $\mathrm{Beta}(1, J)$ distribution we have, for all $t_\star$ large enough, that

$$\begin{aligned}\mathbb{P}\Big(\Big|\varphi_{i,j} - \frac{1}{2}\Big| \le \frac{2}{\log^3(t_\star)} \,\Big|\, T_j\Big) &= J \int_{\frac{1}{2}-2/\log^3(t_\star)}^{\frac{1}{2}+2/\log^3(t_\star)} (1-x)^{J-1}\, dx \\ &\le J\Big(\frac{2}{3}\Big)^{J-1} \frac{4}{\log^3(t_\star)} \le \frac{6}{\log^3(t_\star)},\end{aligned}$$

where we used that $J(2/3)^{J-1} \le 4/3$ for every positive integer J. Taking an expectation over T_j we obtain that

$$\mathbb{P}\Big(\Big|\varphi_{i,j} - \frac{1}{2}\Big| \le \frac{2}{\log^3(t_\star)}\Big) \le \frac{6}{\log^3(t_\star)} \tag{A.13}$$

for all $t_\star$ large enough. We can now bound the probability of $\mathcal{E}_3(i, j)^c$:

$$\mathbb{P}(\mathcal{E}_3(i,j)^c) \le \mathbb{P}\Big(\Big|\varphi_{i,j} - \frac{1}{2}\Big| \le \frac{2}{\log^3(t_\star)}\Big) + \mathbb{P}\Big(\mathcal{E}_3(i,j)^c \cap \Big\{\Big|\varphi_{i,j} - \frac{1}{2}\Big| > \frac{2}{\log^3(t_\star)}\Big\}\Big).$$

By (A.13) the first term above is at most $6/\log^3(t_\star)$ for all $t_\star$ large enough, so what remains is to bound the second term. To do this, we introduce the event

$$\mathcal{E}' := \Big\{\exists t \ge t_\star : \Big|\frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} - \varphi_{i,j}\Big| \ge \frac{1}{\log^3(t_\star)}\Big\}.$$

By the triangle inequality and a union bound we have that

$$\begin{aligned}&\mathbb{P}\Big(\mathcal{E}_3(i,j)^c \cap \Big\{\Big|\varphi_{i,j} - \frac{1}{2}\Big| > \frac{2}{\log^3(t_\star)}\Big\}\Big) \\ &\quad\le \mathbb{P}(\mathcal{E}_2(i,j)^c) + \mathbb{P}(\mathcal{E}_2(j,i)^c) + \mathbb{P}(\mathcal{E}' \cap \mathcal{E}_2(i,j) \cap \mathcal{E}_2(i,j)).\end{aligned}$$

The first two terms in the display above are bounded above by $C/\log^3(t_\star)$ for some finite C, by Step 2. It thus remains to bound the third term. To do this, we condition on the tree $T_{t_\star}$. By the tower rule, noting that $\mathcal{E}_2(i, j)$ and $\mathcal{E}_2(i, j)$ are measurable with respect to $T_{t_\star}$, we have that

$$\mathbb{P}(\mathcal{E}' \cap \mathcal{E}_2(i,j) \cap \mathcal{E}_2(i,j)) = \mathbb{E}\big[\mathbb{P}(\mathcal{E}' \,|\, T_{t_\star})\mathbf{1}_{\mathcal{E}_2(i,j)\cap\mathcal{E}_2(j,i)}\big]. \tag{A.14}$$

Now if $\mathcal{E}'$ holds then there exists $t \ge t_\star$ such that

$$\Big|\frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} - \frac{|(T_{t_\star}, v_i)_{v_j\downarrow}|}{|(T_{t_\star}, v_i)_{v_j\downarrow}| + |(T_{t_\star}, v_j)_{v_i\downarrow}|}\Big| \ge \frac{1}{2\log^3(t_\star)}.$$

Therefore, by Lemma A.3, we have that

$$\mathbb{P}(\mathcal{E}' \,|\, T_{t_\star}) \le 2\exp\Big(-\frac{|(T_{t_\star}, v_i)_{v_j\downarrow}| + |(T_{t_\star}, v_j)_{v_i\downarrow}|}{8\log^6(t_\star)}\Big).$$

By (A.10) this implies that

$$\mathbb{P}(\mathcal{E}' \,|\, T_{t_\star})\mathbf{1}_{\mathcal{E}_2(i,j)\cap\mathcal{E}_2(j,i)} \le 2\exp\Big(-\frac{1}{8}t_\star \log^{-13}(t_\star)\Big)$$

and so by (A.14) we have that

$$\mathbb{P}(\mathcal{E}' \cap \mathcal{E}_2(i,j) \cap \mathcal{E}_2(i,j)) \le 2\exp\Big(-\frac{1}{8}t_\star \log^{-13}(t_\star)\Big).$$

Putting everything together we have thus shown for UA trees that

$$\mathbb{P}(\mathcal{E}_3(i,j)^c) \le \frac{C}{\log^3(t_\star)}$$

for some finite constant C and all $t_\star \ge 2$.

The proof for PA trees is similar, so we only highlight the minor changes. First, conditioned on T_j, the pair

$$\big(2|(T_t, v_i)_{v_j\downarrow}| - 1, 2|(T_t, v_j)_{v_i\downarrow}| - 1\big),$$

when viewed at times when one of the coordinates increases, evolves as a Pólya urn with replacement matrix $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$, started from $(1, 2J-1)$. This implies that, conditioned on T_j, we have that $\varphi_{i,j} \sim \text{Beta}(1/2, J - 1/2)$. The probability estimate with the beta distribution follows similarly, resulting in the inequality in (A.13), with the constant 6 replaced with a larger finite constant. The rest of the proof is unchanged, except when Lemma A.3 is applied, then the constant in the exponent changes.

We have thus shown, for both PA and UA trees, that

$$\mathbb{P}(\mathcal{A}_3^c) \le \sum_{\substack{1\le i,j\le \log t_\star \\ i\ne j}} \mathbb{P}(\mathcal{A}_3(i,j)^c) \le \sum_{\substack{1\le i,j\le \log t_\star \\ i\ne j}} \mathbb{P}(\mathcal{E}_3(i,j)^c) \le \sum_{\substack{1\le i,j\le \log t_\star \\ i\ne j}} \frac{C}{\log^3(t_\star)} \le \frac{C}{\log t_\star}$$

for some finite constant C and all $t_\star \ge 2$.

Brief recap. We briefly pause to recap what we have proved so far. Observe that on the event $\mathcal{A}_1 \cap \mathcal{A}_3$ we have that property (A1) of Definition 4.3 holds. In Steps 1 and 3 above we proved that $\mathbb{P}((\mathcal{A}_1 \cap \mathcal{A}_3)^c) \le \mathbb{P}(\mathcal{A}_1^c) + \mathbb{P}(\mathcal{A}_3^c) \le C/\log t_\star$ for some finite constant C and all $t_\star \ge 2$. What remains is to deal with properties (A2) and (A3) of Definition 4.3.

Step 4: *The root of the largest pendent subtree of the centroid is an early vertex.*

Recall the definition of $\widetilde{v}_{i,t}(1)$ from Section 4.1: $\widetilde{v}_{i,t}(1)$ is the neighbor of v_i that is the root of the largest subtree of (T_t, v_i) (assuming that there is a unique largest subtree; if the largest subtree is not unique, let $\widetilde{v}_{i,t}(1)$ denote a neighbor of v_i that is the root of a largest subtree of (T_t, v_i)). For $i \le \log t_\star$, define the event

$$\mathcal{A}_4(i) := \{\forall t \ge t_\star \text{ the timestamp of } \widetilde{v}_{i,t}(1) \text{ is at most } \log t_\star\}.$$

Since $\widetilde{v}_{i,t}(1)$ may not be uniquely defined, the definition of $\mathcal{A}_4(i)$ needs some clarification: in the definition of $\mathcal{A}_4(i)$ it is understood that, if $\widetilde{v}_{i,t}(1)$ is not uniquely defined, then *every* vertex that can be chosen as $\widetilde{v}_{i,t}(1)$ has timestamp at most $\log t_\star$. In other words, $\mathcal{A}_4(i)$ is the event that no neighbor of v_i with timestamp greater than $\log t_\star$ is the root of a largest subtree of (T_t, v_i), for all $t \ge t_\star$. Define also

$$\mathcal{A}_4 := \bigcap_{1\le i\le \log t_\star} \mathcal{A}_4(i).$$

Our goal in Step 4 is to bound $\mathbb{P}(\mathcal{A}_4^c)$.

To abbreviate notation, in the following we let $s := \log t_\star$ and fix $i \le s$. For any $t \ge s$ we define two subtrees. First, let $T'_t(i) := (T_t, v_i)_{\widetilde{v}_{i,s}(1)\downarrow}$; here if $\widetilde{v}_{i,s}(1)$ is not uniquely defined, then we fix a particular choice for the remainder of the argument. We also define $T''_t(i)$ to be the subtree of T_t rooted at v_i that contains all subtrees of (T_t, v_i) formed after time s. In particular, we have that $|T'_s(i)| = \Psi_s(v_i)$ and $|T''_s(i)| = 1$. Now define the event

$$\mathcal{E}_4(i) := \left\{\forall t \ge s : \frac{|T''_t(i)|}{|T''_t(i)| + |T'_t(i)|} < \frac{1}{2}\right\}.$$

If $\mathcal{E}_4(i)$ holds, then $|T''_t(i)| < |T'_t(i)|$ for all $t \ge t_\star$, which implies that no subtree of v_i born after time s ever becomes as large as the subtree rooted at $\widetilde{v}_{i,s}(1)$. Therefore, if $\mathcal{E}_4(i)$ holds,

then $\mathcal{A}_4(i)$ must also hold. Thus $\mathbb{P}(\mathcal{A}_4(i)^c) \le \mathbb{P}(\mathcal{E}_4(i)^c)$, and in the following we bound this latter probability.

Consider first the case of UA trees. Conditioned on T_s, the pair $(|T_t''(i)|, |T_t'(i)|)$, when viewed at times when one of the coordinates increases, evolves as a classical Pólya urn started from $(1, \Psi_s(v_i))$. Therefore, Lemma A.3 implies that

$$\mathbb{P}\left(\exists t \ge s : \frac{|T_t''(i)|}{|T_t''(i)| + |T_t'(i)|} \ge \frac{1}{1+\Psi_s(v_i)} + \lambda \,\middle|\, T_s\right) \le \exp\left(-\frac{\lambda^2}{2}\Psi_s(v_i)\right) \tag{A.15}$$

for every $\lambda > 0$. For PA trees a similar argument shows that (A.15) holds with a different constant in the exponent, and for all $\lambda \ge 2/\Psi_s(v_i)$.

Recalling that $d_n(v)$ denotes the degree of v in T_n, define the event

$$\mathcal{E}'(i) := \{d_s(v_i) < \sqrt{s}\log^2(s)\}.$$

By Lemma A.4 we have, for both PA and UA trees, that

$$\mathbb{P}(\mathcal{E}'(i)^c) \le \exp(-c\log^3(s)) = \exp(-c(\log\log t_\star)^3) \tag{A.16}$$

for some positive constant c. On the event $\mathcal{E}'(i)$ we have that

$$\Psi_s(v_i) = |(T_s, v_i)_{\tilde{v}_{i,s}(1)\downarrow}| \ge \frac{s-1}{\sqrt{s}\log^2(s)} \ge \log^{1/3}(t_\star),$$

where the second inequality holds for all $t_\star$ large enough. Here the first inequality follows from the pigeonhole principle: there are $s-1$ vertices in the rooted subtree (T_s, v_i) apart from v_i, and there are at most $\sqrt{s}\log^2(s)$ subtrees, so at least one of them has at least $(s-1)/(\sqrt{s}\log^2(s))$ vertices.

Combining this argument with the inequality (A.15), we have, for all $t_\star$ large enough, that

$$\mathbb{P}(\mathcal{E}_4(i)^c \mid T_s)\mathbf{1}_{\mathcal{E}'(i)} \le \exp(-c\log^{1/3}(t_\star)) \tag{A.17}$$

for some positive constant c, and both PA and UA trees. Putting together (A.16) and (A.17) we thus have that

$$\begin{aligned}\mathbb{P}(\mathcal{E}_4(i)^c) &= \mathbb{E}[\mathbb{P}(\mathcal{E}_4(i)^c \mid T_s)] \le \mathbb{E}[\mathbb{P}(\mathcal{E}_4(i)^c \mid T_s)\mathbf{1}_{\mathcal{E}'(i)}] + \mathbb{P}(\mathcal{E}'(i)^c) \\ &\le \exp(-c\log^{1/3}(t_\star)) + \exp(-c(\log\log t_\star)^3) \le 2\exp(-c(\log\log t_\star)^3)\end{aligned}$$

for some positive constant c and all $t_\star$ large enough. Finally, by a union bound we have that

$$\mathbb{P}(\mathcal{A}_4^c) \le \sum_{i=1}^{\log t_\star} \mathbb{P}(\mathcal{A}_4(i)^c) \le \sum_{i=1}^{\log t_\star} \mathbb{P}(\mathcal{E}_4(i)^c) \le 2\log(t_\star)\exp(-c(\log\log t_\star)^3)$$

for some positive constant c and all $t_\star$ large enough. This is at most $1/\log t_\star$ for all $t_\star$ large enough.

Step 5: *Early subtree rankings are stable.*

For i satisfying $1 \le i \le \log t_\star$, let $\mathcal{A}_5(i)$ denote the event that for every pair of neighbors u_1, u_2 of v_i that are early vertices (that is, have timestamp at most $\log t_\star$), we either have that $|(T_t, v_i)_{u_1\downarrow}| > |(T_t, v_i)_{u_2\downarrow}|$ for all $t \ge t_\star$ or that $|(T_t, v_i)_{u_1\downarrow}| < |(T_t, v_i)_{u_2\downarrow}|$ for all $t \ge t_\star$. In other words, the pairwise rankings of early subtrees of v_i do not change after time $t_\star$. Define also $\mathcal{A}_5 := \bigcap_{1\le i\le \log t_\star} \mathcal{A}_5(i)$.

Observe that, since u_1 and u_2 are neighbors of v_i, we have that $(T_t, v_i)_{u_1\downarrow} = (T_t, u_2)_{u_1\downarrow}$ and that $(T_t, v_i)_{u_2\downarrow} = (T_t, u_1)_{u_2\downarrow}$. Let $k, \ell \le \log t_\star$ be distinct positive integers and recall from Step 3 that, for all $t_\star$ large enough, on the event $\mathcal{E}_3(k,\ell)$ we either have that $|(T_t, v_k)_{v_\ell\downarrow}| >$

$|(T_t, v_\ell)_{v_k\downarrow}|$ for all $t \geq t_\star$ or that $|(T_t, v_k)_{v_\ell\downarrow}| < |(T_t, v_\ell)_{v_k\downarrow}|$ for all $t \geq t_\star$. Putting the previous two sentences together we have that

$$\bigcap_{\substack{1\leq i,j\leq \log t_\star \\ i\neq j}} \mathcal{E}_3(i,j) \subseteq \mathcal{A}_5$$

for all $t_\star$ large enough. Consequently, by Step 3 we have, for some finite constant C and all $t_\star$ large enough, that

$$\mathbb{P}(\mathcal{A}_5^c) \leq \sum_{\substack{1\leq i,j\leq \log t_\star \\ i\neq j}} \mathbb{P}(\mathcal{E}_3(i,j)^c) \leq \sum_{\substack{1\leq i,j\leq \log t_\star \\ i\neq j}} \frac{C}{\log^3(t_\star)} \leq \frac{C}{\log t_\star}.$$

Finally, observe that on the event $\mathcal{A}_1 \cap \mathcal{A}_4 \cap \mathcal{A}_5$ we have that property (A2) of Definition 4.3 holds. Furthermore, we have shown that $\mathbb{P}((\mathcal{A}_1 \cap \mathcal{A}_4 \cap \mathcal{A}_5)^c) \leq \mathbb{P}(\mathcal{A}_1^c) + \mathbb{P}(\mathcal{A}_4^c) + \mathbb{P}(\mathcal{A}_5^c) \leq C/\log t_\star$ for some finite constant C and all $t_\star \geq 2$.

Step 6: *Concentration for early subtrees.*

It remains to deal with property (A3) of Definition 4.3. So far we have shown that on the event $\mathcal{A}_1 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5$ we have that properties (A1) and (A2) of Definition 4.3 hold, and moreover that $\theta(t_\star)$ and $\widetilde{\theta}_{t_\star}(1)$ are both early vertices. In light of this we define the events

$$\mathcal{A}_6(i,j) := \left\{\forall t \geq t_\star : \left|\frac{1}{t}\right|(T_t, v_i)_{v_j\downarrow}\Big| - \left|\frac{1}{t_\star}\right|(T_{t_\star}, v_i)_{v_j\downarrow}|| \leq \frac{1}{t_\star^{1/3}} \cdot \frac{1}{t_\star}|(T_{t_\star}, v_i)_{v_j\downarrow}|\right\}$$

for distinct positive integers $i, j \leq \log t_\star$, and also

$$\mathcal{A}_6 := \bigcap_{\substack{1\leq i,j\leq \log t_\star \\ i\neq j}} \mathcal{A}_6(i,j).$$

Observe that on the event $\mathcal{A}_1 \cap \mathcal{A}_3 \cap \mathcal{A}_4 \cap \mathcal{A}_5 \cap \mathcal{A}_6$ we have that property (A3) of Definition 4.3 holds. In fact, since property (A3) lies at the crux of the proof, this is the reason why we introduced the events $\mathcal{A}_1, \ldots, \mathcal{A}_6$. Note that while property (A3) concerns the anti-centrality of the centroid, on the event $\mathcal{A}_1$ the centroid is an early vertex, and on the event $\mathcal{A}_4$ the root of the largest pendent subtree of the centroid is an early vertex, which explains why the intersection in the definition of $\mathcal{A}_6$ goes over early pairs of indices. Thus to conclude the proof what remains to be shown is that $\mathbb{P}(\mathcal{A}_6^c) \leq C/\log t_\star$ for some finite constant C and all $t_\star \geq 2$.

Fix distinct positive integers $i, j \leq \log t_\star$. By arguments similar to those in Step 3, in particular using Lemma A.3, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{A}_6(i,j)^c \mid T_{t_\star}) &\leq 2\exp\left(-ct_\star\left(\frac{1}{t_\star^{1/3}} \cdot \frac{1}{t_\star}|(T_{t_\star}, v_i)_{v_j\downarrow}|\right)^2\right) \\ &= 2\exp\left(-ct_\star^{-5/3}|(T_{t_\star}, v_i)_{v_j\downarrow}|^2\right)\end{aligned}$$

for some positive constant c and all $t_\star$ large enough. Recalling the definition of $\mathcal{E}_2(i,j)$ we thus have that

$$\mathbb{P}(\mathcal{A}_6(i,j)^c \mid T_{t_\star})\mathbf{1}_{\mathcal{E}_2(i,j)} \leq 2\exp\left(-ct_\star^{1/3}\log^{-14}(t_\star)\right)$$

for all $t_\star$ large enough. Using Step 2 we thus have that

$$\begin{aligned}\mathbb{P}(\mathcal{A}_6(i,j)^c) &\leq \mathbb{E}\left[\mathbb{P}(\mathcal{A}_6(i,j)^c \mid T_{t_\star})\mathbf{1}_{\mathcal{E}_2(i,j)}\right] + \mathbb{P}(\mathcal{E}_2(i,j)^c) \\ &\leq 2\exp\left(-ct_\star^{1/3}\log^{-14}(t_\star)\right) + \frac{12}{\log^3(t_\star)}\end{aligned}$$

for all $t_\star$ large enough. The conclusion follows by a union bound. □

APPENDIX B: PROOFS OF REMAINING LEMMAS IN SECTION 5

B.1. Proof of Lemma 5.4. We start with a preliminary lemma.

LEMMA B.1. *Let $\{T_n\}_{n\geq 2}$ be a sequence of trees started from the seed $S = S_2$ and grown according to PA or UA. Let $v_1, v_2, v_3, \ldots$ denote the vertices in order of appearance. Let $d_n(v)$ denote the degree of v in T_n. Fix $\varepsilon > 0$. There exists a finite constant C such that the following holds. For every $t_\star \geq C$, $t \geq t_\star^{\varepsilon}$, and $i \leq 100 \log t_\star$, we have that*

$$\mathbb{P}\Big(d_t(v_i) \leq \frac{1}{6}\log t\Big) \leq t^{-1/28}.$$

PROOF. If $i \leq i'$, then $d_t(v_i)$ stochastically dominates $d_t(v_{i'})$ for every $t \geq i'$, so it suffices to prove the inequality for $i = i_\star := \lfloor 100 \log t_\star \rfloor > 2$. Let $t > i_\star$, and let $X_{i_\star+1}, \ldots, X_t$ be independent Bernoulli random variables such that $\mathbb{E}[X_k] = 1/(2k-4)$. Conditioned on T_{t-1}, the probability that v_t connects to $v_{i_\star}$ is at least $1/(2t-4)$, for both PA and UA trees. This implies that $d_t(v_{i_\star})$ stochastically dominates $Y_t := 1 + X_{i_\star+1} + \cdots X_t$. Thus we have that

$$\mathbb{P}\Big(d_t(v_{i_\star}) \leq \frac{1}{6}\log t\Big) \leq \mathbb{P}\Big(Y_t \leq \frac{1}{6}\log t\Big).$$

Since $Y_t - 1$ is the sum of independent Bernoulli random variables, we have that $\mathrm{Var}(Y_t) \leq \mathbb{E}[Y_t]$. Thus by Bernstein's inequality we have for every $x \geq 0$ that

$$\mathbb{P}\big(Y_t \leq \mathbb{E}[Y_t] - x\big) \leq \exp\Big(-\frac{x^2/2}{\mathbb{E}[Y_t] + x/3}\Big).$$

Setting $x = \mathbb{E}[Y_t]/2$, we obtain that

$$\mathbb{P}\Big(Y_t \leq \frac{1}{2}\mathbb{E}[Y_t]\Big) \leq \exp\Big(-\frac{3}{28}\mathbb{E}[Y_t]\Big).$$

We have that

$$\mathbb{E}[Y_t] = 1 + \frac{1}{2}\sum_{k=i_\star-1}^{t-2}\frac{1}{k} \geq 1 + \frac{1}{2}\int_{i_\star-1}^{t-1}\frac{1}{x}\,dx = 1 + \frac{1}{2}\log\Big(\frac{t-1}{i_\star-1}\Big)$$

and so $\mathbb{E}[Y_t] \geq \frac{1}{3}\log t$ for all $t_\star$ large enough. Plugging this inequality into the displays above and putting them together concludes the proof. □

PROOF OF LEMMA 5.4. The proof is similar to that of Lemma 4.4 and most of the work has already been done there. However, we modify the proof in a few key places to show the desired result. First, we slightly change the definition of an early vertex. Fix $\gamma := 18$. In the following we informally call a vertex an "early" vertex if its timestamp is at most $\gamma \log t_\star$. We also fix $s_1 := t_\star^{1/64}$ and $s_2 := t_\star^{1/48}$, and note that $s_1 = s_2^{3/4}$.

Modified Step 1: *The centroid is an early vertex.*

Recall the definition of $\mathcal{A}_1(i)$ from Lemma 4.4. Define $\mathcal{C}_1 := \bigcap_{i>\gamma\log t_\star} \mathcal{A}_1(i)$, the event that only early vertices are ever a centroid. Similarly as in Step 1 of Lemma 4.4, we thus have, for all $t_\star$ large enough, that

$$\mathbb{P}(\mathcal{C}_1^c) \leq \sum_{i>\gamma\log t_\star} \mathbb{P}(\mathcal{A}_1(i)^c) \leq \sum_{i>\gamma\log t_\star} e^{-i/3} \leq \frac{4}{t_\star^{\gamma/3}}. \tag{B.1}$$

Now let $\mathcal{D}_1$ denote the event that $T_{t_\star}$ satisfies $\mathbb{P}(\mathcal{C}_1^c \mid T_{t_\star}) \le t_\star^{-\gamma/6}$. By Markov's inequality, the tower rule, and (B.1) we have that

$$\mathbb{P}(\mathcal{D}_1^c) = \mathbb{P}\big(\mathbb{P}(\mathcal{C}_1^c \mid T_{t_\star}) > t_\star^{-\gamma/6}\big) \le t_\star^{\gamma/6}\mathbb{E}\big[\mathbb{P}(\mathcal{C}_1^c \mid T_{t_\star})\big] = t_\star^{\gamma/6}\mathbb{P}(\mathcal{C}_1^c) \le 4t_\star^{-\gamma/6}, \tag{B.2}$$

where the last inequality holds for all $t_\star$ large enough.

Modified Step 2: *Subtrees formed before time s_2 are large in $T_{t_\star}$.*

In Step 2 of Lemma 4.4 we proved that early subtrees are large in $T_{t_\star}$. Here we need to show that many more subtrees are large—though what "large" means is relaxed here—for reasons that will become clear in later steps. Formally, define the $T_{t_\star}$-measurable events

$$\mathcal{D}_2(i, j) := \{|(T_{t_\star}, v_i)_{v_j\downarrow}| \ge t_\star^{7/8}\}$$

for distinct positive integers $i, j \le s_2$, and also

$$\mathcal{D}_2 := \bigcap_{\substack{1\le i,j\le s_2\\ i\ne j}} \mathcal{D}_2(i, j).$$

We proceed by bounding the probability of the complement of $\mathcal{D}_2(i, j)$. Since the arguments are identical to those in Step 2 of Lemma 4.4, we omit most details and only give the final bounds.

Assume in the following that $1 \le i < j \le s_2$. In Step 2 of Lemma 4.4 we showed that, for both PA and UA trees, and for every $z \in [0, 1]$, we have that

$$\max\left\{\mathbb{P}\left(\frac{1}{t_\star}|(T_{t_\star}, v_i)_{v_j\downarrow}| \le z\right), \mathbb{P}\left(\frac{1}{t_\star}|(T_{t_\star}, v_j)_{v_i\downarrow}| \le z\right)\right\} \le 12\sqrt{jz}.$$

Setting $z = t_\star^{-1/8}$ and using the bound $j \le s_2$, we obtain that

$$\max\{\mathbb{P}(\mathcal{D}_2(i, j)^c), \mathbb{P}(\mathcal{D}_2(i, j)^c)\} \le 12s_2^{1/2}t_\star^{-1/16}.$$

By a union bound we thus have, for both PA and UA trees, that

$$\mathbb{P}(\mathcal{D}_2^c) \le 12s_2^{5/2}t_\star^{-1/16} \le 12t_\star^{-1/96}. \tag{B.3}$$

Modified Step 3: *The size-based ranking in $T_{t_\star}$ of subtrees formed before time s_2 persists.*

This is similar to Step 3 of Lemma 4.4, but with some differences, which we highlight. Define the events

$$\mathcal{C}_3(i, j) := \{\forall t \ge t_\star : \big(|(T_{t_\star}, v_i)_{v_j\downarrow}| - |(T_{t_\star}, v_j)_{v_i\downarrow}|\big)\big(|(T_t, v_i)_{v_j\downarrow}| - |(T_t, v_j)_{v_i\downarrow}|\big) > 0\},$$

$$\mathcal{H}_3(i, j) := \left\{\forall t \ge t_\star : \left|\frac{|(T_t, v_i)_{v_j\downarrow}|}{|(T_t, v_i)_{v_j\downarrow}| + |(T_t, v_j)_{v_i\downarrow}|} - \frac{1}{2}\right| > t_\star^{-1/4}\right\},$$

$$\mathcal{D}_3(i, j) := \left\{\left|\frac{|(T_{t_\star}, v_i)_{v_j\downarrow}|}{|(T_{t_\star}, v_i)_{v_j\downarrow}| + |(T_{t_\star}, v_j)_{v_i\downarrow}|} - \frac{1}{2}\right| \ge 2t_\star^{-1/4}\right\}$$

for distinct positive integers $i, j \le s_2$, and also

$$\mathcal{C}_3 := \bigcap_{\substack{1\le i,j\le s_2\\ i\ne j}} \mathcal{C}_3(i, j) \quad \text{and} \quad \mathcal{D}_3 := \bigcap_{\substack{1\le i,j\le s_2\\ i\ne j}} \mathcal{D}_3(i, j).$$

By the same arguments as in Step 3 of Lemma 4.4, we have that if $\mathcal{H}_3(i, j) \cap \mathcal{D}_2$ holds, then $\mathcal{C}_3(i, j)$ must also hold. By Lemma A.3 we have, for every tree $T_{t_\star}$ such that $\mathcal{D}_3(i, j) \cap \mathcal{D}_2$ holds, that

$$\mathbb{P}(\mathcal{H}_3(i, j)^c \mid T_{t_\star}) \le 2\exp\big(-(2t_\star^{7/8} - 1)t_\star^{-1/2}/8\big) \le 2\exp\big(-t_\star^{3/8}/8\big).$$

Thus by a union bound we have, for every tree $T_{t_\star}$ such that $\mathcal{D}_3 \cap \mathcal{D}_2$ holds, that

$$\mathbb{P}(\mathcal{C}_3^c \,|\, T_{t_\star}) \le 2s_2^2 \exp(-t_\star^{3/8}/8), \tag{B.4}$$

and note that this decays faster than any polynomial in $t_\star$.

In the remainder of this step we bound the probability $\mathbb{P}((\mathcal{D}_3 \cap \mathcal{D}_2)^c)$. A union bound shows that $\mathbb{P}((\mathcal{D}_3 \cap \mathcal{D}_2)^c) \le \mathbb{P}(\mathcal{D}_2^c) + \mathbb{P}(\mathcal{D}_3^c \cap \mathcal{D}_2)$; the former probability is at most $12t_\star^{-1/96}$ by (B.3), so it suffices to bound $\mathbb{P}(\mathcal{D}_3^c \cap \mathcal{D}_2)$. By a further union bound, and incurring a factor of s_2^2, it suffices to bound $\mathbb{P}(\mathcal{D}_3(i,j)^c \cap \mathcal{D}_2)$, where $1 \le i < j \le s_2$. To this end, define $\varphi_{i,j}$ as in (A.12); again this limiting random variable exists almost surely. By a union bound we then have that

$$\begin{aligned}&\mathbb{P}(\mathcal{D}_3(i,j)^c \cap \mathcal{D}_2)\\ &\quad\le \mathbb{P}\Big(\Big|\varphi_{i,j} - \frac{1}{2}\Big| \le 4t_\star^{-1/4}\Big) + \mathbb{P}\Big(\mathcal{D}_3(i,j)^c \cap \mathcal{D}_2 \cap \Big\{\Big|\varphi_{i,j} - \frac{1}{2}\Big| > 4t_\star^{-1/4}\Big\}\Big).\end{aligned}$$

Both of these terms can be bounded by the same arguments as in Step 3 of Lemma 4.4. First, there exists a finite absolute constant C such that the first term above is at most $Ct_\star^{-1/4}$. Next, the second term is at most $2\exp(-t_\star^{3/8}/8)$. Altogether this gives that $\mathbb{P}(\mathcal{D}_3(i,j)^c \cap \mathcal{D}_2) \le C' t_\star^{-1/4}$ for some finite absolute constant C'. By a union bound we thus have that $\mathbb{P}(\mathcal{D}_3^c \cap \mathcal{D}_2) \le C' t_\star^{-5/24}$. Putting everything together we have thus obtained that

$$\mathbb{P}((\mathcal{D}_3 \cap \mathcal{D}_2)^c) \le C'' t_\star^{-1/96} \tag{B.5}$$

for some finite absolute constant C''.

Modified Step 4: *The roots of the K largest pendent subtrees of the centroid have timestamp at most s_2.*

This is similar to Step 4 of Lemma 4.4, but with significant differences—this step has the biggest differences among all. For one, we have to additionally show that the centroid has degree at least K.

For a positive integer $i \le \gamma \log t_\star$ define the event

$$\mathcal{D}_4'(i) := \Big\{ d_{s_1}(v_i) > \frac{1}{6}\log s_1 \Big\},$$

and also define $\mathcal{D}_4' := \bigcap_{1 \le i \le \gamma \log t_\star} \mathcal{D}_4'(i)$. By Lemma B.1 and a union bound we have that

$$\mathbb{P}((\mathcal{D}_4')^c) \le \sum_{i=1}^{\gamma \log t_\star} \mathbb{P}((\mathcal{D}_4'(i))^c) \le (\gamma \log t_\star) s_1^{-1/28} = (\gamma \log t_\star) t_\star^{-1/1792} \tag{B.6}$$

for all $t_\star$ large enough. Observe that if T_{s_1} is such that $\mathcal{D}_4'$ holds, then—since $K \le (1/6)\log s_1$—all early vertices have degree at least K in T_{s_1}, and hence also in T_t for $t \ge s_1$ (in particular $t = t_\star$). For every T_{s_1} such that $\mathcal{D}_4'$ holds, and for every $i \le \gamma \log t_\star$, choose and fix K neighbors of v_i in T_{s_1} arbitrarily (e.g., the K neighbors with largest pendent subtrees: $\widetilde{v}_{i,s_1}(1), \dots, \widetilde{v}_{i,s_1}(K)$, with ties broken by favoring earlier vertices), and label them as $u_1^i, \dots, u_K^i$. In the following, whenever we refer to a tree T_{s_1} such that $\mathcal{D}_4'$ holds, we automatically assume this fixed choice of $K\gamma \log t_\star$ labeled vertices (where repetitions are possible). In the following we fix T_{s_1} such that $\mathcal{D}_4'$ holds and condition on T_{s_1}.

Now fix $i \le \gamma \log t_\star$. To simplify notation, we write $u_1, \dots, u_K$ instead of $u_1^i, \dots, u_K^i$. By Pólya urn arguments it follows that (conditioned on T_{s_1}) for every $\ell \in [K]$ the limiting random variable

$$\phi_{i,\ell} := \lim_{t\to\infty} \frac{1}{t} |(T_t, v_i)_{u_\ell \downarrow}|$$

exists almost surely. Moreover, its distribution (conditioned on T_{s_1}) is given by

$$\phi_{i,\ell} \sim \begin{cases} \text{Beta}\big(|(T_{s_1}, v_i)_{u_\ell\downarrow}|, s_1 - |(T_{s_1}, v_i)_{u_\ell\downarrow}|\big) & \text{for UA,} \\ \text{Beta}\Big(|(T_{s_1}, v_i)_{u_\ell\downarrow}| - \frac{1}{2}, s_1 - |(T_{s_1}, v_i)_{u_\ell\downarrow}| - \frac{1}{2}\Big) & \text{for PA.} \end{cases}$$

We now argue that, for both PA and UA trees, for all $x \in [0, 1]$ we have that

$$\mathbb{P}(\phi_{i,\ell} < x \mid T_{s_1}) \le 2\sqrt{s_1 x}. \tag{B.7}$$

We start with UA trees. When $|(T_{s_1}, v_i)_{u_\ell\downarrow}| = 1$, a direct computation shows that

$$\mathbb{P}(\phi_{i,\ell} < x \mid T_{s_1}) = 1 - (1-x)^{s_1 - 1} \le s_1 x.$$

Otherwise, Markov's inequality implies that

$$\mathbb{P}(\phi_{i,\ell} < x \mid T_{s_1}) = \mathbb{P}\big(\phi_{i,\ell}^{-1} > x^{-1} \mid T_{s_1}\big) \le x\mathbb{E}\big[\phi_{i,\ell}^{-1} \mid T_{s_1}\big] = x\frac{s_1 - 1}{|(T_{s_1}, v_i)_{u_\ell\downarrow}| - 1} \le s_1 x. \tag{B.8}$$

For PA trees, when $|(T_{s_1}, v_i)_{u_\ell\downarrow}| = 1$, a direct computation shows that

$$\mathbb{P}(\phi_{i,\ell} < x \mid T_{s_1}) = \frac{1}{B(\frac{1}{2}, s_1 - \frac{3}{2})} \int_0^x y^{-1/2}(1-y)^{s_1 - 5/2}\, dy \le \sqrt{s_1} \int_0^x y^{-1/2}\, dy = 2\sqrt{s_1 x},$$

where in the inequality we used that $B(\frac{1}{2}, s_1 - \frac{3}{2}) \ge 1/\sqrt{s_1}$ (see (A.9)) and also that $t_\star$ is large enough (so that $s_1 \ge 5/2$). Otherwise, Markov's inequality (just like in (B.8)) implies a bound of $2s_1 x$. In conclusion, we have shown (B.7) in all cases. As a consequence, using the same martingale arguments as in the proof of Lemma 4.6, we have that

$$\mathbb{P}\Big(\frac{1}{s_2}|(T_{s_2}, v_i)_{u_\ell\downarrow}| \le x \,\Big|\, T_{s_1}\Big) \le 2\mathbb{P}(\phi_{i,\ell} \le 4x \mid T_{s_1}) \le 8\sqrt{s_1 x}. \tag{B.9}$$

Now define the event

$$\mathcal{D}_4''(i) := \bigcap_{1 \le \ell \le K} \big\{|(T_{s_2}, v_i)_{u_\ell\downarrow}| \ge s_2^{1/8}\big\},$$

which is well defined when T_{s_1} is such that $\mathcal{D}_4'$ holds. By a union bound and using (B.9) with $x = s_2^{-7/8}$, we have that

$$\mathbb{P}\big(\mathcal{D}_4''(i)^c \mid T_{s_1}\big) \le \sum_{\ell=1}^{K} \mathbb{P}\big(|(T_{s_2}, v_i)_{u_\ell\downarrow}| < s_2^{1/8} \mid T_{s_1}\big) \le 8K s_2^{-1/16} = 8K t_\star^{-1/768}.$$

Now define the event $\mathcal{D}_4'' := \bigcap_{1 \le i \le \gamma \log t_\star} \mathcal{D}_4''(i)$, which is well defined when T_{s_1} is such that $\mathcal{D}_4'$ holds. By the display above, together with a union bound, we have, for every T_{s_1} such that $\mathcal{D}_4'$ holds, that

$$\mathbb{P}\big((\mathcal{D}_4'')^c \mid T_{s_1}\big) \le (8K\gamma \log t_\star) t_\star^{-1/768}. \tag{B.10}$$

For $i \le \gamma \log t_\star$ define the event

$$\begin{aligned} \mathcal{C}_4(i) := \big\{&d_{t_\star}(v_i) \ge K\big\} \\ &\cap \big\{\forall t \ge t_\star : \text{the timestamps of } \widetilde{v}_{i,t}(1), \dots, \widetilde{v}_{i,t}(K) \text{are all at most } s_2\big\}, \end{aligned}$$

and also let $\mathcal{C}_4 := \bigcap_{1 \le i \le \gamma \log t_\star} \mathcal{C}_4(i)$. Note that if $\mathcal{D}_4'$ holds, then $\{d_{t_\star}(v_i) \ge K\}$ holds as well, so to understand $\mathcal{C}_4(i)$ we need to understand the second event in the display above. To do

this, we consider the subtree T_t' of (T_t, v_i) which is rooted at v_i and consists of v_i together with all subtrees of v_i that are formed after time s_2. We can then define the event

$$\mathcal{H}_4(i) := \bigcap_{1\le \ell \le K} \left\{ \forall t \ge s_2 : \left| \frac{|T_t'|}{|T_t'| + |(T_t, v_i)_{u_\ell\downarrow}|} - \frac{1}{1 + |(T_{s_2}, v_i)_{u_\ell\downarrow}|} \right| \le \frac{1}{3} \right\},$$

which is well defined whenever T_{s_1} is such that $\mathcal{D}_4'$ holds. Provided that $t_\star$ is large enough, if $\mathcal{H}_4(i)$ holds, then $|T_t'|/(|T_t'| + |(T_t, v_i)_{u_\ell\downarrow}|) < 1/2$ for all $t \ge s_2$, which implies that no subtree born after time s_2 will ever become larger than any of the subtrees with roots $u_1, \dots, u_K$. This, in turn, means that no subtree born after time s_2 will ever become one of the K largest subtrees of v_i. Therefore, $\mathcal{H}_4(i) \subseteq \mathcal{C}_4(i)$.

If T_{s_1} is such that $\mathcal{D}_4'$ holds, and also T_{s_2} is such that $\mathcal{D}_4''$ holds, then by Lemma A.3 and a union bound we have that

$$\mathbb{P}(\mathcal{H}_4(i)^c \mid T_{s_1}, T_{s_2}) \le 2K \exp\left(-\frac{1}{72} t_\star^{1/384}\right).$$

Together with the previous paragraph and a union bound we thus have that

$$\mathbb{P}(\mathcal{C}_4^c \mid T_{s_1}, T_{s_2}) \le (2K\gamma \log t_\star) \exp\left(-\frac{1}{72} t_\star^{1/384}\right) \tag{B.11}$$

whenever T_{s_1} is such that $\mathcal{D}_4'$ holds, and also T_{s_2} is such that $\mathcal{D}_4''$ holds.

The display above motivates defining $\mathcal{D}_4$ to be the event that $T_{t_\star}$ satisfies

$$\mathbb{P}(\mathcal{C}_4^c \mid T_{t_\star}) \le \exp\left(-\frac{1}{144} t_\star^{1/384}\right); \tag{B.12}$$

note that $\mathcal{D}_4$ is $T_{t_\star}$-measurable. In the rest of this step we bound $\mathbb{P}(\mathcal{D}_4^c)$. By conditioning first on T_{s_1} and then on T_{s_2}, together with a couple of union bounds, we obtain that

$$\mathbb{P}(\mathcal{D}_4^c) \le \mathbb{E}\big[\mathbb{P}(\mathcal{D}_4^c \mid T_{s_1}, T_{s_2}) \mathbf{1}_{\mathcal{D}_4'} \mathbf{1}_{\mathcal{D}_4''}\big] + \mathbb{E}\big[\mathbb{P}((\mathcal{D}_4'')^c \mid T_{s_1}) \mathbf{1}_{\mathcal{D}_4'}\big] + \mathbb{P}((\mathcal{D}_4')^c). \tag{B.13}$$

By (B.6) and (B.10) we have that the second and the third term in the display above are together at most $t_\star^{-1/1800}$ for all $t_\star$ large enough. Turning to the first term in the display above, let T_{s_1} be such that $\mathcal{D}_4'$ holds, and subsequently let T_{s_2} be such that $\mathcal{D}_4''$ holds. Then by Markov's inequality we have that

$$\begin{aligned}
\mathbb{P}(\mathcal{D}_4^c \mid T_{s_1}, T_{s_2}) &= \mathbb{P}\left(\mathbb{P}(\mathcal{C}_4^c \mid T_{t_\star}) > \exp\left(-\frac{1}{144} t_\star^{1/384}\right) \,\middle|\, T_{s_1}, T_{s_2}\right) \\
&\le \exp\left(\frac{1}{144} t_\star^{1/384}\right) \mathbb{E}\big[\mathbb{P}(\mathcal{C}_4^c \mid T_{t_\star}) \mid T_{s_1}, T_{s_2}\big] \\
&= \exp\left(\frac{1}{144} t_\star^{1/384}\right) \mathbb{P}(\mathcal{C}_4^c \mid T_{s_1}, T_{s_2}).
\end{aligned}$$

Now plugging in (B.11), we obtain that

$$\mathbb{P}(\mathcal{D}_4^c \mid T_{s_1}, T_{s_2}) \le (2K\gamma \log t_\star) \exp\left(-\frac{1}{144} t_\star^{1/384}\right).$$

Plugging this back into (B.13) we finally obtain, for all $t_\star$ large enough, that

$$\mathbb{P}(\mathcal{D}_4^c) \le 2 t_\star^{-1/1800}. \tag{B.14}$$

Modified Step 5: In Step 5 of Lemma 4.4 we showed that early subtree rankings are stable. Here we already showed in Modified Step 3 that the size-based ranking in $T_{t_\star}$ of subtrees formed before time s_2 persists.

Modified Step 6: *Concentration of subtree sizes.*
In light of the previous steps, we define the events

$$\mathcal{C}_6(i,j) := \left\{\forall t \geq t_\star : \left|\frac{1}{t}\right|(T_t, v_i)_{v_j\downarrow}\Big| - \frac{1}{t_\star}\Big|(T_{t_\star}, v_i)_{v_j\downarrow}|| \leq \frac{1}{t_\star^{1/3}} \cdot \frac{1}{t_\star}|(T_{t_\star}, v_i)_{v_j\downarrow}|\right\}$$

for distinct positive integers $i, j \leq s_2$, and also

$$\mathcal{C}_6 := \bigcap_{\substack{1\leq i,j\leq s_2 \\ i\neq j}} \mathcal{C}_6(i,j).$$

In Step 6 of Lemma 4.4 we showed that

$$\mathbb{P}\big(\mathcal{C}_6(i,j)^c \,|\, T_{t_\star}\big) \leq 2\exp\big(-ct_\star^{-5/3}|(T_{t_\star}, v_i)_{v_j\downarrow}|^2\big)$$

for some positive constant c and all $t_\star$ large enough. Thus if $T_{t_\star}$ is such that $\mathcal{D}_2$ holds, then

$$\mathbb{P}\big(\mathcal{C}_6(i,j)^c \,|\, T_{t_\star}\big) \leq 2\exp\big(-ct_\star^{-5/3}t_\star^{7/4}\big) = 2\exp\big(-ct_\star^{1/12}\big).$$

Thus by a union bound we have that if $T_{t_\star}$ is such that $\mathcal{D}_2$ holds, then

$$\mathbb{P}\big(\mathcal{C}_6^c \,|\, T_{t_\star}\big) \leq 2s_2^2\exp\big(-ct_\star^{1/12}\big), \tag{B.15}$$

which decays faster than any polynomial in $t_\star$.

Putting everything together. Define the events

$$\mathcal{D} := \mathcal{D}_1 \cap \mathcal{D}_2 \cap \mathcal{D}_3 \cap \mathcal{D}_4,$$
$$\widetilde{\mathcal{C}} := \mathcal{C}_1 \cap \mathcal{C}_3 \cap \mathcal{C}_4 \cap \mathcal{C}_6.$$

The event $\mathcal{D}$ is $T_{t_\star}$-measurable by construction. Putting together (B.2), (B.5), and (B.14), we have that $\mathbb{P}(\mathcal{D}^c) \leq 3t_\star^{-1/1800}$ for all $t_\star$ large enough.

Next, we argue that if $\mathcal{D}$ holds, then $|T_{t_\star}(k)| \geq t_\star^{7/8}$. First, note that if $\mathcal{D}_1$ holds, then the centroid at time $t_\star$ is an early vertex. If $\mathcal{D}_4$ holds, then all early vertices have degree at least K in $T_{t_\star}$, and for every early vertex the timestamps of their neighbors corresponding to the K largest pendent subtrees are all at most s_2. Finally, if $\mathcal{D}_2$ holds, then all subtrees formed before time s_2 have size at least $t_\star^{7/8}$, and if $\mathcal{D}_3$ holds, then none of these subtree sizes are equal (i.e., everything is well defined). Putting these observations together we indeed have that $|T_{t_\star}(k)| \geq t_\star^{7/8}$ if $\mathcal{D}$ holds.

Finally, turning to the event $\mathcal{C}$, observe that $\widetilde{\mathcal{C}} \subseteq \mathcal{C}$ by construction. Therefore,

$$\mathbb{P}\big(\mathcal{C}^c \,|\, \mathcal{D}\big) \leq \mathbb{P}\big(\widetilde{\mathcal{C}}^c \,|\, \mathcal{D}\big)$$

and it suffices to bound this latter quantity. Putting together the definition of $\mathcal{D}_1$, (B.4), the definition of $\mathcal{D}_4$ (see (B.12)), and (B.15), we have that for every tree $T_{t_\star}$ such that $\mathcal{D}$ holds, we have that

$$\mathbb{P}\big(\widetilde{\mathcal{C}}^c \,|\, T_{t_\star}\big) \leq Ct_\star^{-\gamma/6}$$

for some universal finite constant C. Taking an expectation over $T_{t_\star}$ and recalling that $\gamma = 18$ concludes the proof of (5.4), and thus also the proof of the lemma. □

B.2. Proof of the variance estimate in Lemma 5.6. We start with two preliminary lemmas regarding the variance and covariance of functions of Beta and Dirichlet random variables, which will be useful in the proof of Lemma 5.6.

LEMMA B.2. *There exists a finite constant C such that the following holds. Let α and t be such that $1/2 \le \alpha < t$ and $1/2 \le t - \alpha$. Let ψ_1 and ψ_2 be i.i.d.* Beta$(\alpha, t-\alpha)$ *random variables. Then*

$$\operatorname{Var}\big((\psi_1 - \psi_2)^2\big) \le \frac{C\alpha^2(t-\alpha)^2}{t^6}.$$

PROOF. Let $\psi \sim \text{Beta}(\alpha, t-\alpha)$. Bounding the variance by the second moment we have that

$$\begin{aligned}\operatorname{Var}\big((\psi_1 - \psi_2)^2\big) &\le \mathbb{E}\big[(\psi_1 - \psi_2)^4\big] = \mathbb{E}\big[\psi_1^4 - 4\psi_1^3\psi_2 + 6\psi_1^2\psi_2^2 - 4\psi_1\psi_2^3 + \psi_2^4\big] \\ &= 2\mathbb{E}\big[\psi^4\big] - 8\mathbb{E}\big[\psi^3\big]\mathbb{E}[\psi] + 6\mathbb{E}\big[\psi^2\big]^2.\end{aligned}$$

For every positive integer k we have that $\mathbb{E}[\psi^k] = \prod_{i=0}^{k-1}(\alpha+i)/(t+i)$. Plugging this into the display above we obtain that

$$\mathbb{E}\big[(\psi_1 - \psi_2)^4\big] = \frac{12\alpha(\alpha+1)(t-\alpha)(t-\alpha+1)}{t^2(t+1)^2(t+2)(t+3)}$$

and the claim follows. □

LEMMA B.3. *There exists a finite constant C such that the following holds. Let α_1, α_2, and t be such that $1/2 \le \alpha_1, \alpha_2$ and $\alpha_1 + \alpha_2 < t$. Let $(\psi_1, \phi_1, 1-\psi_1-\phi_1)$ and $(\psi_2, \phi_2, 1-\psi_2-\phi_2)$ be i.i.d.* Dir$(\alpha_1, \alpha_2, t-\alpha_1-\alpha_2)$ *random vectors, where* Dir *denotes the Dirichlet distribution. Then*

$$\operatorname{Cov}\big((\psi_1 - \psi_2)^2, (\phi_1 - \phi_2)^2\big) \le \frac{C\alpha_1^2\alpha_2^2}{t^6}.$$

PROOF. Let $(\psi, \phi, 1-\psi-\phi) \sim \text{Dir}(\alpha_1, \alpha_2, t-\alpha_1-\alpha_2)$. By expanding the terms in the definition of the covariance and using independence, we have that

$$\begin{aligned}&\operatorname{Cov}\big((\psi_1 - \psi_2)^2, (\phi_1 - \phi_2)^2\big) \\ &\quad= 2\mathbb{E}\big[\psi^2\phi^2\big] - 2\mathbb{E}\big[\psi^2\big]\mathbb{E}\big[\phi^2\big] + 4\mathbb{E}[\psi]\big\{\mathbb{E}[\psi]\mathbb{E}\big[\phi^2\big] - \mathbb{E}\big[\psi\phi^2\big]\big\} \\ &\qquad+ 4\mathbb{E}[\phi]\big\{\mathbb{E}\big[\psi^2\big]\mathbb{E}[\phi] - \mathbb{E}\big[\psi^2\phi\big]\big\} + 4\mathbb{E}[\psi\phi]^2 - 4\big(\mathbb{E}[\psi]\mathbb{E}[\phi]\big)^2.\end{aligned}$$

For nonnegative integers β_1 and β_2, the joint moments of ψ and ϕ are given by

$$\mathbb{E}\big[\psi^{\beta_1}\phi^{\beta_2}\big] = \frac{\prod_{i=0}^{\beta_1-1}(\alpha_1+i)\prod_{j=0}^{\beta_2-1}(\alpha_2+j)}{\prod_{i=0}^{\beta_1+\beta_2-1}(t+i)}.$$

Plugging this into the display above we obtain that

$$\begin{aligned}&\operatorname{Cov}\big((\psi_1 - \psi_2)^2, (\phi_1 - \phi_2)^2\big) \\ &= \frac{4\alpha_1\alpha_2\{-2t^3 + (2\alpha_1\alpha_2 + 5\alpha_1 + 5\alpha_2 - 3)t^2 + (-5\alpha_1\alpha_2 + 6\alpha_1 + 6\alpha_2)t - 6\alpha_1\alpha_2\}}{t^4(t+1)^2(t+2)(t+3)}.\end{aligned}$$

To obtain an upper bound, we can drop all negative terms in the numerator. Using also the trivial bounds $t \le t^2$ and $\alpha_1, \alpha_2 \le 2\alpha_1\alpha_2$, we thus obtain that

$$\operatorname{Cov}\big((\psi_1 - \psi_2)^2, (\phi_1 - \phi_2)^2\big) \le \frac{200\alpha_1^2\alpha_2^2}{t^2(t+1)^2(t+2)(t+3)}$$

and the claim follows. □

We are now ready to prove Lemma 5.6.

PROOF OF LEMMA 5.6. We bound the variance by conditioning on the tree $T_{t_\star}$. By the law of total variance we have that

$$\begin{aligned}\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,\mathcal{D}\big) &= \mathbb{E}\big[\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big)\,\big|\,\mathcal{D}\big]\\ &\quad+\mathbb{E}\big[\mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big]^2\,\big|\,\mathcal{D}\big]-\mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,\mathcal{D}\big]^2.\end{aligned}$$

From the proof of Lemma 5.5 (see also the proof of Lemma 4.7) it follows that

$$\limsup_{n\to\infty}\mathbb{E}\big[\mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big]^2\,\big|\,\mathcal{D}\big]\le\left(\frac{1+3t_\star^{-1/3}}{t_\star}\right)^2$$

for all $t_\star$ large enough, and by Lemma 5.5 we also have that

$$\liminf_{n\to\infty}\mathbb{E}\big[S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,\mathcal{D}\big]^2\ge\left(\frac{1-3t_\star^{-1/3}}{t_\star}\right)^2$$

for all $t_\star$ large enough (where in both cases "large enough" does not depend on k). Putting these displays together we obtain that

$$\limsup_{n\to\infty}\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,\mathcal{D}\big)\le\limsup_{n\to\infty}\mathbb{E}\big[\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big)\,\big|\,\mathcal{D}\big]+\frac{12}{t_\star^{7/3}}$$

for all $t_\star$ large enough. Since $k\le K\le\log t_\star$, the latter term in the display above is at most $12/(kt_\star^2)$, so it remains to bound the first term.

Interchanging the limsup and the expectation, we have that

$$\limsup_{n\to\infty}\mathbb{E}\big[\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big)\,\big|\,\mathcal{D}\big]\le\mathbb{E}\Big[\limsup_{n\to\infty}\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big)\,\big|\,\mathcal{D}\Big],\tag{B.16}$$

so in what follows we study the conditional variance of $S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}$ given $T_{t_\star}$ (with $T_{t_\star}$ such that $\mathcal{D}$ holds). Expanding the variance of the sum we have that

$$\operatorname{Var}\big(S_n(k)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big)=\frac{1}{k^2}\sum_{\ell=1}^{k}\sum_{m=1}^{k}\operatorname{Cov}\big(Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2},Y_n(m)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big).\tag{B.17}$$

Recall from Section 5.1 the definition of $Z_n^i(\ell)$, the limit $Z^i(\ell):=\lim_{n\to\infty}Z_n^i(\ell)$, and the distribution of the limit from (5.3). In particular, recall that on the event $\mathcal{C}^1\cap\mathcal{C}^2$ we have that $X_n^i(\ell)=Z_n^i(\ell)$ for all $n\ge t_\star$ and all $1\le\ell\le K$. To bound the covariance in (B.17), we bound from above the expectation of the product, and bound from below the individual expectations. First, using property (C3) of Definition 5.2 we have that

$$\begin{aligned}&\mathbb{E}\big[Y_n(\ell)Y_n(m)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big]\\ &\quad\le\frac{\mathbb{E}[(X_n^1(\ell)-X_n^2(\ell))^2(X_n^1(m)-X_n^2(m))^2\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,|\,T_{t_\star}]}{4(1-t_\star^{-1/3})^4\frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})}\\ &\quad\le\frac{\mathbb{E}[(Z_n^1(\ell)-Z_n^2(\ell))^2(Z_n^1(m)-Z_n^2(m))^2\,|\,T_{t_\star}]}{4(1-t_\star^{-1/3})^4\frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})},\end{aligned}\tag{B.18}$$

where the second inequality follows by replacing $X_n^i(\ell)$ and $X_n^i(m)$ with $Z_n^i(\ell)$ and $Z_n^i(m)$ on the event $\mathcal{C}^1\cap\mathcal{C}^2$, and then removing the indicator. Turning to the lower bound, from the proof of Lemma 5.5 we have, for any $\ell\le K$ and any $T_{t_\star}$ such that $\mathcal{D}$ holds, that

$$\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}\,\big|\,T_{t_\star}\big]\ge\frac{\mathbb{E}[(Z_n^1(\ell)-Z_n^2(\ell))^2\,|\,T_{t_\star}]}{2(1+t_\star^{-1/3})^2\frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})}-t_\star^{1/8}\mathbb{P}\big((\mathcal{C}^1\cap\mathcal{C}^2)^c\,\big|\,T_{t_\star}\big).$$

On the event $\mathcal{D}$ we have that $|T_{t_\star}(\ell)| \geq t_\star^{7/8}$, which implies that the fraction in the display above is at most $t_\star^{1/8}$. Therefore, multiplying the bounds in the display above with indices ℓ and m we obtain that

$$\begin{aligned}
&\mathbb{E}\big[Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,\big|\, T_{t_\star}\big]\mathbb{E}\big[Y_n(m)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,\big|\, T_{t_\star}\big] \\
&\quad\geq \frac{\mathbb{E}[(Z_n^1(\ell)-Z_n^2(\ell))^2 \,|\, T_{t_\star}]\mathbb{E}[(Z_n^1(m)-Z_n^2(m))^2 \,|\, T_{t_\star}]}{4(1+t_\star^{-1/3})^4 \frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})} \\
&\quad- 2t_\star^{1/4}\mathbb{P}\big((\mathcal{C}^1\cap\mathcal{C}^2)^c \,\big|\, T_{t_\star}\big).
\end{aligned} \tag{B.19}$$

Putting together (B.18) and (B.19), we obtain an upper bound on the covariance in (B.17) that consists of three terms:

$$\begin{aligned}
&\mathrm{Cov}\big(Y_n(\ell)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2}, Y_n(m)\mathbf{1}_{\mathcal{C}^1\cap\mathcal{C}^2} \,\big|\, T_{t_\star}\big) \\
&\quad\leq \frac{\mathrm{Cov}((Z_n^1(\ell)-Z_n^2(\ell))^2, (Z_n^1(m)-Z_n^2(m))^2 \,|\, T_{t_\star})}{4(1-t_\star^{-1/3})^4 \frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})} \\
&\quad+ \big\{(1-t_\star^{-1/3})^{-4} - (1+t_\star^{-1/3})^{-4}\big\}\times \\
&\quad\times \frac{\mathbb{E}[(Z_n^1(\ell)-Z_n^2(\ell))^2 \,|\, T_{t_\star}]\mathbb{E}[(Z_n^1(m)-Z_n^2(m))^2 \,|\, T_{t_\star}]}{4\frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})} \\
&\quad+ 2t_\star^{1/4}\mathbb{P}\big((\mathcal{C}^1\cap\mathcal{C}^2)^c \,\big|\, T_{t_\star}\big).
\end{aligned} \tag{B.20}$$

We now deal with each term in turn, starting with the last one. Since this term does not depend on the indices ℓ and m, nor on n, averaging over ℓ and m, and taking the limit as $n\to\infty$, this term remains $2t_\star^{1/4}\mathbb{P}((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, T_{t_\star})$. Taking an expectation over $T_{t_\star}$ (see (B.16)), this becomes $2t_\star^{1/4}\mathbb{P}((\mathcal{C}^1\cap\mathcal{C}^2)^c \,|\, \mathcal{D})$, which by Lemma 5.4 is at most $C/t_\star^{11/4}$ for some finite constant C.

Turning to the second term in (B.20), first note that

$$(1-t_\star^{-1/3})^{-4} - (1+t_\star^{-1/3})^{-4} \leq 9t_\star^{-1/3}$$

for all $t_\star$ large enough. In the proof of Lemma 5.5 we showed that

$$\lim_{n\to\infty} \frac{\mathbb{E}[(Z_n^1(\ell)-Z_n^2(\ell))^2 \,|\, T_{t_\star}]}{\frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})} \leq \frac{C}{t_\star}$$

for all $\ell\leq K$ and some universal finite constant C. Putting these bounds together, we obtain that, after taking a limit as $n\to\infty$ (which exists), the second term in (B.20) is at most $C/t_\star^{7/3}$ for some universal finite constant C. This holds for all indices ℓ and m, and for all trees $T_{t_\star}$. Thus after averaging over all these we still have a bound of $C/t_\star^{7/3}$.

Finally, we turn to the first term in (B.20), which is the main term among the three. By the bounded convergence theorem the limit as $n\to\infty$ of this term exists and is equal to

$$\frac{\mathrm{Cov}((Z^1(\ell)-Z^2(\ell))^2, (Z^1(m)-Z^2(m))^2 \,|\, T_{t_\star})}{4(1-t_\star^{-1/3})^4 \frac{|T_{t_\star}(\ell)|}{t_\star}(1-\frac{|T_{t_\star}(\ell)|}{t_\star})\frac{|T_{t_\star}(m)|}{t_\star}(1-\frac{|T_{t_\star}(m)|}{t_\star})}.$$

To obtain a slightly simpler expression, recall that $|T_{t_\star}(\ell)|\leq t_\star/2$ for all $\ell\in\{1,\dots,K\}$, and hence the display above is bounded from above by

$$\frac{Ct_\star^2\,\mathrm{Cov}((Z^1(\ell)-Z^2(\ell))^2, (Z^1(m)-Z^2(m))^2 \,|\, T_{t_\star})}{|T_{t_\star}(\ell)||T_{t_\star}(m)|} \tag{B.21}$$

for some universal finite constant C. We now distinguish two cases based on whether or not the indices ℓ and m are equal.

First, when $\ell = m$, we have from (5.3) and Lemma B.2 that

$$\operatorname{Var}\big((Z^1(\ell) - Z^2(\ell))^2 \,|\, T_{t_\star}\big) \le \frac{C|T_{t_\star}(\ell)|^2(t_\star - |T_{t_\star}(\ell)|)^2}{t_\star^6} \le \frac{C|T_{t_\star}(\ell)|^2}{t_\star^4}$$

for some universal finite constant C. Thus the expression in (B.21) is bounded from above by $C'/t_\star^2$ for some universal finite constant C'. There are k terms in (B.17) where the indices are equal; furthermore, there is a $1/k^2$ factor in front of the sum. Putting all this together we see that the contribution from these terms is at most $C'/(kt_\star^2)$, which is the bound in the claim.

We turn now to the case when $\ell \neq m$. By Pólya urn arguments (see, e.g., [42], Section 4.5) it follows that the two vectors $(Z^1(\ell), Z^1(m), 1 - Z^1(\ell) - Z^1(m))$ and $(Z^2(\ell), Z^2(m), 1 - Z^2(\ell) - Z^2(m))$ are i.i.d. (conditionally given $T_{t_\star}$) Dirichlet random vectors, with parameters given as follows:

$$\begin{aligned}&\big(Z(\ell), Z(m), 1 - Z(\ell) - Z(m)\big)\\ &\quad\sim \begin{cases} \operatorname{Dir}\big(|T_{t_\star}(\ell)|, |T_{t_\star}(m)|, t_\star - |T_{t_\star}(\ell)| - |T_{t_\star}(m)|\big) & \text{for UA,}\\ \operatorname{Dir}\Big(|T_{t_\star}(\ell)| - \frac{1}{2}, |T_{t_\star}(m)| - \frac{1}{2}, t_\star - |T_{t_\star}(\ell)| - |T_{t_\star}(m)|\Big) & \text{for PA.}\end{cases}\end{aligned}$$

By Lemma B.3 we thus have for $\ell \neq m$ that

$$\operatorname{Cov}\big((Z^1(\ell) - Z^2(\ell))^2, (Z^1(m) - Z^2(m))^2 \,|\, T_{t_\star}\big) \le \frac{C|T_{t_\star}(\ell)|^2|T_{t_\star}(m)|^2}{t_\star^6}$$

for some universal finite constant C. Thus the expression in (B.21) is bounded from above by $C'|T_{t_\star}(\ell)||T_{t_\star}(m)|/t_\star^4$ for some universal finite constant C'. Plugging this back into (B.17) we see that the contribution to this expression from terms where $\ell \neq m$ is at most

$$\frac{C'}{k^2 t_\star^4}\sum_{\ell=1}^{k}\sum_{m=1}^{k}|T_{t_\star}(\ell)||T_{t_\star}(m)| = \frac{C'}{k^2 t_\star^4}\Bigg(\sum_{\ell=1}^{k}|T_{t_\star}(\ell)|\Bigg)^2 \le \frac{C'}{k^2 t_\star^4}t_\star^2 = \frac{C'}{k^2 t_\star^2},$$

which concludes the claim. □

Acknowledgments. We thank the anonymous reviewers and an anonymous Associate Editor for helpful comments and feedback which helped improve the manuscript.

Funding. The research of M.Z.R. was supported in part by NSF Grant DMS 1811724 and by a Princeton SEAS Innovation Award.

The research of A.S. was supported in part by NSF Grant DMS 1811724.

SUPPLEMENTARY MATERIAL

Detecting correlation in correlated PA and UA trees (DOI: 10.1214/21-AAP1703SUPP; .pdf). In this supplementary material we give alternative proofs to Corollaries 1.4 and 1.5 that are algorithmic: they explicitly specify (efficiently computable) statistics that detect correlation in PA and UA trees.

REFERENCES

[1] BANERJEE, S., BHAMIDI, S. and CARMICHAEL, I. (2018). Fluctuation bounds for continuous time branching processes and nonparametric change point detection in growing networks. Preprint. Available at https://arxiv.org/abs/1808.02439.

[2] BARABÁSI, A.-L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286** 509–512. MR2091634 https://doi.org/10.1126/science.286.5439.509

[3] BARAK, B., CHOU, C.-N., LEI, Z., SCHRAMM, T. and SHENG, Y. (2019). (Nearly) efficient algorithms for the graph matching problem on correlated random graphs. In *Advances in Neural Information Processing Systems* (*NeurIPS*) 9190–9198.

[4] BERG, A. C., BERG, T. L. and MALIK, J. (2005). Shape matching and object recognition using low distortion correspondences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 26–33.

[5] BERGER, N., BORGS, C., CHAYES, J. T. and SABERI, A. (2014). Asymptotic behavior and distributional limits of preferential attachment graphs. *Ann. Probab.* **42** 1–40. MR3161480 https://doi.org/10.1214/12-AOP755

[6] BHAMIDI, S., JIN, J. and NOBEL, A. (2018). Change point detection in network models: Preferential attachment and long range dependence. *Ann. Appl. Probab.* **28** 35–78. MR3770872 https://doi.org/10.1214/17-AAP1297

[7] BOLLOBÁS, B., RIORDAN, O., SPENCER, J. and TUSNÁDY, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures Algorithms* **18** 279–290. MR1824277 https://doi.org/10.1002/rsa.1009

[8] BUBECK, S., DEVROYE, L. and LUGOSI, G. (2017). Finding Adam in random growing trees. *Random Structures Algorithms* **50** 158–172. MR3607120 https://doi.org/10.1002/rsa.20649

[9] BUBECK, S., ELDAN, R., MOSSEL, E. and RÁCZ, M. Z. (2017). From trees to seeds: On the inference of the seed from large trees in the uniform attachment model. *Bernoulli* **23** 2887–2916. MR3648049 https://doi.org/10.3150/16-BEJ831

[10] BUBECK, S., MOSSEL, E. and RÁCZ, M. Z. (2015). On the influence of the seed graph in the preferential attachment model. *IEEE Trans. Netw. Sci. Eng.* **2** 30–39. MR3361606 https://doi.org/10.1109/TNSE.2015.2397592

[11] CHO, M. and LEE, K. M. (2012). Progressive graph matching: Making a move of graphs via probabilistic voting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*) 398–405. IEEE, New York.

[12] CONTE, D., FOGGIA, P., SANSONE, C. and VENTO, M. (2004). Thirty years of graph matching in pattern recognition. *Int. J. Pattern Recognit. Artif. Intell.* **18** 265–298.

[13] COUR, T., SRINIVASAN, P. and SHI, J. (2007). Balanced graph matching. In *Advances in Neural Information Processing Systems* 313–320.

[14] CULLINA, D. and KIYAVASH, N. (2016). Improved achievability and converse bounds for Erdős-Rényi graph matching. In *ACM SIGMETRICS* **44** 63–72.

[15] CULLINA, D. and KIYAVASH, N. (2018). Exact alignment recovery for correlated Erdős-Rényi graphs. Preprint. Available at https://arxiv.org/abs/1711.06783.

[16] CURIEN, N., DUQUESNE, T., KORTCHEMSKI, I. and MANOLESCU, I. (2015). Scaling limits and influence of the seed graph in preferential attachment trees. *J. Éc. Polytech. Math.* **2** 1–34. MR3326003 https://doi.org/10.5802/jep.15

[17] DEVROYE, L. and REDDAD, T. (2019). On the discovery of the seed in uniform attachment trees. *Internet Math.* 1–29. MR3934290

[18] DING, J., MA, Z., WU, Y. and XU, J. (2021). Efficient random graph matching via degree profiles. *Probab. Theory Related Fields* **179** 29–115. MR4221654 https://doi.org/10.1007/s00440-020-00997-4

[19] DRMOTA, M. (2009). *Random Trees*: *An Interplay Between Combinatorics and Probability*. Springer, Vienna. MR2484382 https://doi.org/10.1007/978-3-211-75357-6

[20] FAN, Z., MAO, C., WU, Y. and XU, J. (2019). Spectral Graph Matching and Regularized Quadratic Relaxations I: The Gaussian Model. Preprint. Available at https://arxiv.org/abs/1907.08880.

[21] FAN, Z., MAO, C., WU, Y. and XU, J. (2019). Spectral Graph Matching and Regularized Quadratic Relaxations II: Erdős-Rényi Graphs and Universality. Preprint. Available at https://arxiv.org/abs/1907.08883.

[22] FANTI, G., KAIROUZ, P., OH, S., RAMCHANDRAN, K. and VISWANATH, P. (2016). Rumor source obfuscation on irregular trees. In *ACM SIGMETRICS* **44** 153–164. https://doi.org/10.1145/2964791.2901471

[23] FANTI, G., KAIROUZ, P., OH, S., RAMCHANDRAN, K. and VISWANATH, P. (2017). Hiding the rumor source. *IEEE Trans. Inf. Theory* **63** 6679–6713. MR3707563 https://doi.org/10.1109/TIT.2017.2696960

[24] FANTI, G., KAIROUZ, P., OH, S. and VISWANATH, P. (2015). Spy vs. spy: Rumor source obfuscation. In *ACM SIGMETRICS* **43** 271–284.

[25] GANASSALI, L. and MASSOULIÉ, L. (2020). From tree matching to sparse graph alignment. In *Conference on Learning Theory* 1633–1665. PMLR.

[26] JOG, V. and LOH, P.-L. (2017). Analysis of centrality in sublinear preferential attachment trees via the Crump-Mode-Jagers branching process. *IEEE Trans. Netw. Sci. Eng.* **4** 1–12. MR3625951 https://doi.org/10.1109/TNSE.2016.2622923

[27] JOG, V. and LOH, P.-L. (2018). Persistence of centrality in random growing trees. *Random Structures Algorithms* **52** 136–157. MR3731614 https://doi.org/10.1002/rsa.20726

[28] KAZEMI, E., HASSANI, S. H. and GROSSGLAUSER, M. (2015). Growing a graph matching from a handful of seeds. *Proc. VLDB Endow.* **8** 1010–1021.

[29] KAZEMI, E., YARTSEVA, L. and GROSSGLAUSER, M. (2015). When can two unlabeled networks be aligned under partial overlap? In *Proceedings of the* 53*rd Annual Allerton Conference on Communication, Control, and Computing* (*Allerton*) 33–42. IEEE, New York.

[30] KORULA, N. and LATTANZI, S. (2014). An efficient reconciliation algorithm for social networks. *Proc. VLDB Endow.* **7** 377–388.

[31] LIVI, L. and RIZZI, A. (2013). The graph matching problem. *PAA Pattern Anal. Appl.* **16** 253–283. MR3084902 https://doi.org/10.1007/s10044-012-0284-8

[32] LUGOSI, G. and PEREIRA, A. S. (2019). Finding the seed of uniform attachment trees. *Electron. J. Probab.* **24** Paper No. 18, 15. MR3925458 https://doi.org/10.1214/19-EJP268

[33] LYZINSKI, V., FISHKIND, D. E. and PRIEBE, C. E. (2014). Seeded graph matching for correlated Erdős-Rényi graphs. *J. Mach. Learn. Res.* **15** 3513–3540. MR3291408

[34] MAHMOUD, H. M. (1992). Distances in random plane-oriented recursive trees. *J. Comput. Appl. Math.* **41** 237–245. MR1181723 https://doi.org/10.1016/0377-0427(92)90252-S

[35] MÓRI, T. F. (2005). The maximum degree of the Barabási-Albert random tree. *Combin. Probab. Comput.* **14** 339–348. MR2138118 https://doi.org/10.1017/S0963548304006133

[36] MOSSEL, E. and XU, J. (2019). Seeded graph matching via large neighborhood statistics. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms* 1005–1014. SIAM, Philadelphia, PA. MR3909531 https://doi.org/10.1137/1.9781611975482.62

[37] NARAYANAN, A. and SHMATIKOV, V. (2009). De-anonymizing social networks. In *Proceedings of the* 30*th IEEE Symposium on Security and Privacy* 173–187. IEEE Comput. Soc., Los Alamitos.

[38] NAVLAKHA, S. and KINGSFORD, C. (2011). Network archaeology: Uncovering ancient networks from present-day interactions. *PLoS Comput. Biol.* **7** e1001119, 16. MR2805381 https://doi.org/10.1371/journal.pcbi.1001119

[39] O'DONNELL, R., WRIGHT, J., WU, C. and ZHOU, Y. (2014). Hardness of robust graph isomorphism, Lasserre gaps, and asymmetry of random graphs. In *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms* 1659–1677. ACM, New York. MR3376480 https://doi.org/10.1137/1.9781611973402.120

[40] OVERGOOR, J., BENSON, A. and UGANDER, J. (2019). Choosing to grow a graph: Modeling network formation as discrete choice. In *Proceedings of the World Wide Web Conference* (*WWW*) 1409–1420. ACM, New York.

[41] PEDARSANI, P. and GROSSGLAUSER, M. (2011). On the privacy of anonymized networks. In *Proceedings of the* 17*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (*KDD*) 1235–1243.

[42] RÁCZ, M. Z. and BUBECK, S. (2017). Basic models and questions in statistical network analysis. *Stat. Surv.* **11** 1–47. MR3696007 https://doi.org/10.1214/17-SS117

[43] RÁCZ, M. Z. and SRIDHAR, A. (2022). Supplement to "Correlated randomly growing graphs." https://doi.org/10.1214/21-AAP1703SUPP

[44] REDNER, S. (1998). How popular is your paper? An empirical study of the citation distribution. *Eur. Phys. J. B* **4** 131–134.

[45] SHAH, D. and ZAMAN, T. (2010). Detecting sources of computer viruses in networks: Theory and experiment. In *ACM SIGMETRICS* **38** 203–214.

[46] SHAH, D. and ZAMAN, T. (2011). Rumors in a network: Who's the culprit? *IEEE Trans. Inf. Theory* **57** 5163–5181. MR2849111 https://doi.org/10.1109/TIT.2011.2158885

[47] SHAH, D. and ZAMAN, T. (2016). Finding rumor sources on random trees. *Oper. Res.* **64** 736–755. MR3515208 https://doi.org/10.1287/opre.2015.1455

[48] SINGH, R., XU, J. and BERGER, B. (2008). Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proc. Natl. Acad. Sci. USA* **105** 12763–12768.

[49] YARTSEVA, L. and GROSSGLAUSER, M. (2013). On the performance of percolation graph matching. In *Proceedings of the First ACM Conference on Online Social Networks* 119–130.