On the Influence of the Seed Graph in the Preferential Attachment Model

Sébastien Bubeck, Elchanan Mossel, and Miklós Z. Rácz

Abstract—We study the influence of the seed graph in the preferential attachment model, focusing on the case of trees. We first show that the seed has no effect from a weak local limit point of view. On the other hand, we conjecture that different seeds lead to different distributions of limiting trees from a total variation point of view. We take a first step in proving this conjecture by showing that seeds with different degree profiles lead to different limiting distributions for the (appropriately normalized) maximum degree, implying that such seeds lead to different (in total variation) limiting trees.

Index Terms—Random trees, preferential attachment, seed graph

1 Introduction

We are interested in the following question: suppose we generate a large graph according to the linear preferential attachment model—can we say anything about the initial (seed) graph? A precise answer to this question could lead to new insights for the diverse applications of the preferential attachment model. In this paper we initiate the theoretical study of the seed's influence. Experimental evidence of the seed's influence already exists in the literature, see, e.g., [16]. For sake of simplicity we focus on *trees* grown according to linear preferential attachment.

For a tree T denote by $d_T(u)$ the degree of vertex u in T, $\Delta(T)$ the maximum degree in T, and $\vec{d}(T) \in \mathbb{N}^{\mathbb{N}}$ the vector of degrees arranged in decreasing order.[1] We refer to $\vec{d}(T)$ as the degree profile of T. For $n \geq k \geq 2$ and a tree T on k vertices we define the random tree $\mathrm{PA}(n,T)$ by induction. First $\mathrm{PA}(k,T) = T$. Then, given $\mathrm{PA}(n,T)$, $\mathrm{PA}(n+1,T)$ is formed from $\mathrm{PA}(n,T)$ by adding a new vertex u and a new edge uv where v is selected at random among vertices in $\mathrm{PA}(n,T)$ according to the following probability distribution:

$$\mathbb{P}(v = i \mid \mathrm{PA}(n,T)) = \frac{d_{\mathrm{PA}(n,T)}(i)}{2(n-1)}.$$

This model was introduced in [10] under the name *Random Plane-Oriented Recursive Trees* but we use here the modern terminology of Preferential Attachment graphs, see [2], [5]. In the following we also denote by S_k the k-vertex star, i.e., the tree where a central vertex is connected to all $k-1$ other vertices.

We want to understand whether there is a relation between T and $\mathrm{PA}(n,T)$ when n becomes very large. We investigate three ways to make this question more formal. They correspond to three different points of view on the limiting tree obtained by letting n go to infinity.

The least refined point of view is to consider the tree $\mathrm{PA}(\infty,T)$ defined on a countable set of vertices that one obtains by continuing the preferential attachment process indefinitely. As observed in [9], in this case the seed does not have any influence: indeed for any tree T, almost surely, $\mathrm{PA}(\infty,T)$ will be the unique isomorphism type of tree with countably many vertices and in which each vertex has infinite degree. In fact this statement holds for *any* model where the degree of each fixed vertex diverges to infinity as the tree grows. For example, this notion of limit does not allow to distinguish between linear and non-linear preferential attachment models, as long as the degree of each fixed node diverges to infinity.

Next we consider the much more subtle and fine-grained notion of a weak local limit introduced in [3]. This notion of graph limits contains information about local neighborhoods of a typical vertex (see Section 4 for a precise definition), and is more powerful than the one considered in the previous paragraph as it can, e.g., distinguish between models having different limiting degree distributions. The weak local limit of the preferential attachment graph was first studied in the case of trees in [15] using branching process techniques, and then later in general in [4] using Pólya urn representations. These papers show that $\mathrm{PA}(n,S_2)$ tends to the so-called Pólya-point graph in the weak local limit sense, and our first theorem utilizes this result to obtain the same for an arbitrary seed:

Theorem 1. *For any tree T the weak local limit of $\mathrm{PA}(n,T)$ is the Pólya-point graph described in [4] with $m = 1$.*

This result says that "locally" (in the Benjamini-Schramm sense) the seed has no effect. The intuitive reason for this result is that in the preferential attachment model most nodes are far from the seed graph and therefore it is expected

1. We artificially continue the vector of degrees with zeros after the $|T|$th coordinate to put all degree profiles on the same space.

- *S. Bubeck is with Princeton University, Princeton, NJ. E-mail: sbubeck@princeton.edu.*
- *E. Mossel is with the University of Pennsylvania and the University of California, Berkeley, CA 94703. E-mail: mossel@stat.berkeley.edu.*
- *M.Z. Racz is with the University of California, Berkeley, CA 94703. E-mail: racz@stat.berkeley.edu.*

Manuscript received 21 Sept. 2014; revised 15 Dec. 2014; accepted 15 Jan. 2015. Date of publication 27 Jan. 2015; date of current version 24 Apr. 2015. Recommended for acceptance by E. Kolaczyk.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TNSE.2015.2397592

Fig. 1. Two trees with six vertices and $\vec{d}(S) = \vec{d}(T)$.

that their neighborhoods will not reveal any information about it.

Finally, we consider the most refined point of view, which we believe to be the most natural one for this problem as well as the richest one (both mathematically and in terms of insights for potential applications). First we rephrase our main question in the terminology of hypothesis testing. Given two potential seed trees T and S, and an observation R which is a tree on n vertices, one wishes to test whether $R \sim \mathrm{PA}(n, T)$ or $R \sim \mathrm{PA}(n, S)$. Our original question then boils down to whether one can design a test with asymptotically (in n) non-negligible power. This is equivalent to studying the total variation distance between $\mathrm{PA}(n, T)$ and $\mathrm{PA}(n, S)$, where recall that the total variation distance between two random variables X and Y taking values in a finite space $\mathcal{X}$ with laws μ and ν is defined as $\mathrm{TV}(X, Y) = \frac{1}{2}\sum_{x\in\mathcal{X}} |\mu(x) - \nu(x)|$. Thus we naturally define

$$\delta(S, T) = \lim_{n\to\infty} \mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T)),$$

where $\mathrm{PA}(n, S)$ and $\mathrm{PA}(n, T)$ are random elements in the finite space of unlabeled trees with n vertices. This limit is well-defined because $\mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T))$ is non-increasing in n (since if $\mathrm{PA}(n, S) = \mathrm{PA}(n, T)$, then the evolution of the random trees can be coupled such that $\mathrm{PA}(n', S) = \mathrm{PA}(n', T)$ for all $n' \geq n$) and always nonnegative. One can propose a test with asymptotically non-negligible power (i.e., a non-trivial test) if and only if $\delta(S, T) > 0$. We believe that in fact this is always the case (except in trivial situations); precisely we make the following conjecture:

Conjecture 1. δ is a metric on isomorphism types of trees with at least three vertices.[2]

In the present work we distinguish trees with different degree profiles.

Theorem 2. *Let S and T be two finite trees on at least three vertices. If $\vec{d}(S) \neq \vec{d}(T)$, then $\delta(S, T) > 0$.*

In fact our proof shows a stronger statement, namely that different degree profiles lead to different limiting distributions for the (appropriately normalized) maximum degree.

The smallest pair of trees that our method cannot as of yet distinguish is depicted in Fig. 1.

In some cases we can say more. For instance, the distance between a fixed tree and a star can be arbitrarily close to 1 if the star is large enough.

Theorem 3. *For any fixed tree T one has*

$$\lim_{k\to\infty} \delta(S_k, T) = 1.$$

2. Clearly δ is a pseudometric on isomorphism types of trees with at least three vertices so the only non-trivial part of the statement is that $\delta(S, T) \neq 0$ for S and T non-isomorphic.

1.1 Follow-Up Work

Following the results and conjectures presented here, in a beautiful work, [7] proved that Conjecture 1 is indeed true. The proof utilizes some of the ideas presented here, in particular by using statistics which are very similar to those we consider in Section 3.2. The proof approach of [7] is much more abstract than ours. By constructing and analyzing a large family of martingales, they are able to show that the limiting distribution of these martingales must differ when starting from two different trees. One of the advantages of the more computational proof presented here is that it allows to more easily derive quantitative bounds for the total variation distance in cases where our results show that the distance is nonzero.

1.2 Organization of the Paper

In the next section we derive results on the limiting distribution of the maximum degree $\Delta(\mathrm{PA}(n, T))$ that are useful in proving Theorems 2 and 3, which we then prove in Section 3.1. In Section 3.2 we describe a particular way of generalizing the notion of maximum degree which we believe should provide a(n alternative) way to prove Conjecture 1. At present we are missing a technical result which we state separately as Conjecture 2 in the same section. The proof of Theorem 1 is in Section 4, while the proof of a key lemma described in Section 2 is presented in Section 5. We conclude the paper with open problems in Section 6.

2 Useful Results on the Maximum Degree

We first recall several results that describe the limiting degree distributions of preferential attachment graphs (Section 2.1), and from these we determine the tail behavior of the maximum degree in Section 2.2, which we then use in the proofs of Theorems 2 and 3. Throughout the paper we label the vertices of $\mathrm{PA}(n, T)$ by $\{1, 2, \ldots, n\}$ in the order in which they are added to the graph, with the vertices of the initial tree labeled in decreasing order of degree, i.e., satisfying $d_T(1) \geq d_T(2) \geq \ldots \geq d_T(|T|)$ (with ties broken arbitrarily). We also define the constant

$$c(a, b) = \frac{\Gamma(2a - 2)}{2^{b-1}\Gamma(a - 1/2)\Gamma(b)}, \tag{1}$$

which will occur multiple times.

2.1 Previous Results

2.1.1 Starting from an Edge

Móri [11] used martingale techniques to study the maximum degree of the preferential attachment tree starting from an edge, and showed that $\Delta(\mathrm{PA}(n, S_2))/\sqrt{n}$ converges almost surely to a random variable which we denote by $D_{\max}(S_2)$. He also showed that for each fixed $i \geq 1$, $d_{\mathrm{PA}(n,S_2)}(i)/\sqrt{n}$ converges almost surely to a random variable which we denote by $D_i(S_2)$, and furthermore that

$D_{\max}(S_2) = \max_{i\geq 1} D_i(S_2)$ almost surely. In light of this, in order to understand $D_{\max}(S_2)$ it is useful to study $\{D_i(S_2)\}_{i\geq 1}$. [11] computes the joint moments of $\{D_i(S_2)\}_{i\geq 1}$; in particular, we have (see [11, Eq. (2.4)]) that for $i \geq 2$,

$$\mathbb{E}D_i(S_2)^r = \frac{\Gamma(i-1)\Gamma(1+r)}{\Gamma\left(i-1+\frac{r}{2}\right)}. \tag{2}$$

Using different methods and slightly different normalization, [13] also study the limiting distribution of $d_{\mathrm{PA}(n,S_2)}(i)$; in particular, they give an explicit expression for the limiting density. Fix $s \geq 1/2$ and define

$$\kappa_s(x) = \Gamma(s)\sqrt{\frac{2}{s\pi}}\exp\left(-\frac{x^2}{2s}\right)U\left(s-1,\frac{1}{2},\frac{x^2}{2s}\right)\mathbf{1}_{\{x>0\}},$$

where $U(a,b,z)$ denotes the confluent hypergeometric function of the second kind, also known as the Kummer U function (see [1, Chapter 13]); it can be shown that this is a density function. Peköz et al. [13] show that for $i \geq 2$ the distributional limit of

$$d_{\mathrm{PA}(n,S_2)}(i)/\left(\mathbb{E}d_{\mathrm{PA}(n,S_2)}(i)^2\right)^{1/2}$$

has density κ_{i-1} (they also give rates of convergence to this limit in the Kolmogorov metric). Let W_s denote a random variable with density κ_s. The moments of W_s (see [13, Section 2]) are given by

$$\mathbb{E}W_s^r = \left(\frac{s}{2}\right)^{r/2}\frac{\Gamma(s)\Gamma(1+r)}{\Gamma\left(s+\frac{r}{2}\right)}, \tag{3}$$

and thus comparing (2) and (3) we see that $D_i(S_2) \stackrel{d}{=} \sqrt{2/(i-1)}W_{i-1}$ for $i \geq 2$.

2.1.2 Starting from an Arbitrary Seed Graph

Since we are interested in the effect of the seed graph, we desire similar results for $\mathrm{PA}(n,T)$ for an arbitrary tree T. One way of viewing $\mathrm{PA}(n,T)$ is to start growing a preferential attachment tree from a single edge and condition on it being T after reaching $|T|$ vertices; $\mathrm{PA}(n,T)$ has the same distribution as $\mathrm{PA}(n,S_2)$ conditioned on $\mathrm{PA}(|T|,S_2) = T$. Due to this the almost sure convergence results of [11] carry over to the setting of an arbitrary seed tree. Thus for every fixed $i \geq 1$, $d_{\mathrm{PA}(n,T)}(i)/\sqrt{n}$ converges almost surely to a random variable which we denote by $D_i(T)$, $\Delta(\mathrm{PA}(n,T))/\sqrt{n}$ converges almost surely to a random variable which we denote by $D_{\max}(T)$, and furthermore $D_{\max}(T) = \max_{i\geq 1} D_i(T)$ almost surely.

In order to understand these limiting distributions, the basic observation is that for any i, $1 \leq i \leq |T|$, $(2(n-1) - d_{\mathrm{PA}(n,T)}(i), d_{\mathrm{PA}(n,T)}(i))$ evolves according to a Pólya urn with replacement matrix $\left(\begin{smallmatrix}2&0\\1&1\end{smallmatrix}\right)$ starting from $(2(|T|-1) - d_T(i), d_T(i))$. Indeed, when a new vertex is added to the tree, either it attaches to vertex i, with probability $d_{\mathrm{PA}(n,T)}(i)/(2n-2)$, in which case both $d_{\mathrm{PA}(n,T)}(i)$ and $2(n-1) - d_{\mathrm{PA}(n,T)}(i)$ increase by one (and hence why the second row of the replacement matrix is $(1\ 1)$), or otherwise it attaches to some other vertex in which case $d_{\mathrm{PA}(n,T)}(i)$ does not increase but $2(n-1) - d_{\mathrm{PA}(n,T)}(i)$ increases by two (and hence why the first row of the replacement matrix is $(2\ 0)$). Janson [8] gives limit theorems for triangular Pólya urns, and also provides information about the limiting distributions; for instance [8, Theorem 1.7] gives a formula for the moments of $D_i(T)$, extending (2) for arbitrary trees T: for every i, $1 \leq i \leq |T|$, we have

$$\mathbb{E}D_i(T)^r = \frac{\Gamma(|T|-1)\Gamma(d_T(i)+r)}{\Gamma(d_T(i))\Gamma\left(|T|-1+\frac{r}{2}\right)}, \tag{4}$$

and for $i > |T|$ we have $\mathbb{E}D_i(T)^r = \Gamma(i-1)\Gamma(1+r)/\Gamma(i-1+r/2)$, just like in (2).

The joint distribution of the limiting degrees in the seed graph, $\left(D_1(T),\ldots,D_{|T|}(T)\right)$, can be understood by viewing the evolution of $\left(d_{\mathrm{PA}(n,T)}(1),\ldots,d_{\mathrm{PA}(n,T)}(|T|)\right)$ in the following way. When adding a new vertex, first decide whether it attaches to one of the initial $|T|$ vertices (with probability $\sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i)/(2n-2)$) or not (with the remaining probability); if it does, then independently pick one of them to attach to with probability proportional to their degrees. In other words, if viewed at times when a new vertex attaches to one of the initial $|T|$ vertices, the joint degree counts of the initial vertices evolve like a standard Pólya urn with $|T|$ colors and identity replacement matrix.

Let $\mathrm{Beta}(a,b)$ denote the beta distribution with parameters a and b (with density proportional to $x^{a-1}(1-x)^{b-1}\mathbf{1}_{\{x\in[0,1]\}}$), let $\mathrm{Dir}(\alpha_1,\ldots\alpha_s)$ denote the Dirichlet distribution with density proportional to $x_1^{\alpha_1-1}\cdots x_s^{\alpha_s-1}\mathbf{1}_{\left\{x\in[0,1]^s,\sum_{i=1}^s x_i=1\right\}}$, and write $X \sim \mathrm{GGa}(a,b)$ for a random variable X having the generalized gamma distribution with density proportional to $x^{a-1}e^{-x^b}\mathbf{1}_{\{x>0\}}$. On the one hand, $(2(n-1) - \sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i), \sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i))$ evolves according to a Pólya urn with replacement matrix $\left(\begin{smallmatrix}2&0\\1&1\end{smallmatrix}\right)$ starting from $(0, 2(|T|-1))$. Janson [8] gives the limiting distribution of $\sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i)/\sqrt{n}$ (see Theorem 1.8 and Example 3.1): $\sum_{i=1}^{|T|} D_i(T) \stackrel{d}{=} 2Z_{|T|}$, where $Z_{|T|} \sim \mathrm{GGa}(2|T|-1,2)$. On the other hand, it is known that in a standard Pólya urn with identity replacement matrix the vector of proportions of each color converges almost surely to a random variable with a Dirichlet distribution with parameters given by the initial counts. These facts, together with the observation in the previous paragraph, lead to the following representation: if X and $Z_{|T|}$ are independent, $X \sim \mathrm{Dir}(d_T(1),\ldots,d_T(|T|))$, and $Z_{|T|} \sim \mathrm{GGa}(2|T|-1,2)$, then

$$\left(D_1(T),\ldots,D_{|T|}(T)\right) \stackrel{d}{=} 2Z_{|T|}X. \tag{5}$$

Recently, [14] gave useful representations for $(D_1(T),\ldots,D_r(T))$ for general r, and the representation above appears as a special case (see [14, Remark 1.9]).

2.2 Tail Behavior

In order to prove Theorem 2 our main tool is to study the tail of the limiting degree distributions. In particular, we use the following key lemma.

Lemma 1. *Let T be a finite tree.*

(a) Let $U \subseteq \{1, 2, \ldots, |T|\}$ be a nonempty subset of the vertices of T, and let $d = \sum_{i \in U} d_T(i)$. Then

$$\mathbb{P}\left(\sum_{i \in U} D_i(T) > t\right) \sim c(|T|, d) t^{1-2|T|+2d} \exp\left(-t^2/4\right) \quad (6)$$

as $t \to \infty$, where the constant c is as in (1).[3]

(b) For every $L > |T|$ there exists a constant $C(L) < \infty$ such that for every $t \geq 1$ we have

$$\sum_{i=L}^{\infty} \mathbb{P}(D_i(T) > t) \leq C(L) t^{3-2L} \exp\left(-t^2/4\right). \quad (7)$$

We postpone the proof of Lemma 1 to Section 5, as it results from a lengthy computation. As an immediate corollary we get the asymptotic tail behavior of $D_{\max}(T)$.

Corollary 1. *Let T be a finite tree and let $m := |\{i \in \{1, \ldots, |T|\} : d_T(i) = \Delta(T)\}|$, where recall that $\Delta(T)$ is the maximum degree in T. Then*

$$\mathbb{P}(D_{\max}(T) > t) \sim m \times c(|T|, \Delta(T)) t^{1-2|T|+2\Delta(T)} \exp\left(-t^2/4\right) \quad (8)$$

as $t \to \infty$, where the constant c is as in (1).

Proof. Recall the fact that $D_{\max}(T) = \max_{i \geq 1} D_i(T)$ almost surely. First, a union bound gives us that

$$\mathbb{P}(D_{\max}(T) > t) \leq \sum_{i=1}^{m} \mathbb{P}(D_i(T) > t) + \sum_{i=m+1}^{|T|} \mathbb{P}(D_i(T) > t) + \sum_{i=|T|+1}^{\infty} \mathbb{P}(D_i(T) > t).$$

Then using Lemma 1 we get the upper bound required for (8): the first sum gives the right hand side of (8), while the other two sums are of smaller order. For the lower bound we first have that

$$\mathbb{P}(D_{\max}(T) > t) \geq \sum_{i=1}^{m} \mathbb{P}(D_i(T) > t) - \sum_{i=1}^{m} \sum_{j=i+1}^{m} \mathbb{P}\left(D_i(T) > t, D_j(T) > t\right). \quad (9)$$

Lemma 1(a) with $U = \{i, j\}$ implies that for any $1 \leq i < j \leq m$,

$$\mathbb{P}\left(D_i(T) > t, D_j(T) > t\right) \leq \mathbb{P}\left(D_i(T) + D_j(T) > 2t\right) \leq C_{i,j}(T) t^{1-2|T|+4\Delta(T)} \exp\left(-t^2\right) \quad (10)$$

for some constant $C_{i,j}(T)$ and all t large enough. The exponent $-t^2$, appearing on the right hand side of (10), is smaller by a constant factor than the exponent $-t^2/4$, appearing in the asymptotic expression for $\mathbb{P}(D_i(T) > t)$ (see (6)). Consequently the second sum on the right hand side of (9) is of smaller order than the first sum, and so we have that $\mathbb{P}(D_{\max}(T) > t) \geq (1 - o(1)) \sum_{i=1}^{m} \mathbb{P}(D_i(T) > t)$ as $t \to \infty$. We conclude using Lemma 1. □

3. Throughout the paper we use standard asymptotic notation; for instance, $f(t) \sim g(t)$ as $t \to \infty$ if $\lim_{t \to \infty} f(t)/g(t) = 1$.

3 Distinguishing Trees Using the Maximum Degree

In this section we first prove Theorems 2 and 3, both using Corollary 1 (see Section 3.1). Then in Section 3.2 we describe a particular way of generalizing the notion of maximum degree which we believe should provide a way to prove Conjecture 1. At present we are missing a technical result, see Conjecture 2 below, and we prove Conjecture 1 assuming that this holds. Although [7] have now proven Conjecture 1, we believe this alternative approach could be of interest by itself due to its simplicity, and it may also lead to better bounds. Moreover, as described at the end of the section, the statistics used by [7] are very similar to the ones we considered, and it would be interesting to understand this connection better.

3.1 Proofs

Proof of Theorem 2. We first provide a simple proof of distinguishing two trees of the same size but with different maximum degree, and then show how to extend this argument to the other cases.

Case 1: $|S| - \Delta(S) \neq |T| - \Delta(T)$. W.l.o.g. suppose that $|S| - \Delta(S) < |T| - \Delta(T)$. Clearly for any $t > 0$ and $n \geq \max\{|S|, |T|\}$ one has

$$\mathrm{TV}(\mathrm{PA}(n, S), \mathrm{PA}(n, T)) \geq \mathrm{TV}(\Delta(\mathrm{PA}(n, S)), \Delta(\mathrm{PA}(n, T))) \geq \mathbb{P}\left(\Delta(\mathrm{PA}(n, S)) > t\sqrt{n}\right) - \mathbb{P}\left(\Delta(\mathrm{PA}(n, T)) > t\sqrt{n}\right).$$

Taking the limit as $n \to \infty$ this implies that

$$\delta(S, T) \geq \sup_{t > 0} \left[\mathbb{P}(D_{\max}(S) > t) - \mathbb{P}(D_{\max}(T) > t)\right]. \quad (11)$$

By Corollary 1 and the fact that $|S| - \Delta(S) < |T| - \Delta(T)$ we have that $\mathbb{P}(D_{\max}(S) > t) > \mathbb{P}(D_{\max}(T) > t)$ for large enough t, which concludes the proof in this case.

Case 2: $|S| \neq |T|$. W.l.o.g. suppose that $|S| < |T|$. If $|S| - \Delta(S) \neq |T| - \Delta(T)$ then by Case 1 we have that $\delta(S, T) > 0$, so we may assume that $|S| - \Delta(S) = |T| - \Delta(T)$. Just as in the proof of Case 1 we have that

$$\delta(S, T) \geq \sup_{t > 0} \left[\mathbb{P}(D_{\max}(T) > t) - \mathbb{P}(D_{\max}(S) > t)\right]. \quad (12)$$

Corollary 1 provides the asymptotic behavior for $\mathbb{P}(D_{\max}(T) > t)$ in the form of (8), where $m \geq 1$.

To find an upper bound for $\mathbb{P}(D_{\max}(S) > t)$, first notice that $\Delta(\mathrm{PA}(|T|, S)) \leq \Delta(T)$, with equality holding if and only if all of the $|T| - |S|$ vertices of $\mathrm{PA}(|T|, S)$ that were added to S connect to the same vertex $i \in \{1, 2, \ldots, |S|\}$ and $d_S(i) = \Delta(S)$. Consequently, if $\Delta(\mathrm{PA}(|T|, S)) = \Delta(T)$, then there is exactly one vertex $j \in \{1, 2, \ldots, |T|\}$ such that $d_{\mathrm{PA}(|T|, S)}(j) = \Delta(T)$. This, together with Corollary 1, shows that on the one hand

$$\mathbb{P}(D_{\max}(S) > t \,|\, \Delta(\mathrm{PA}(|T|, S)) < \Delta(T)) = o(t^{1-2|T|+2\Delta(T)} \exp(-t^2/4)),$$

as $t \to \infty$, and on the other hand

$$\mathbb{P}(D_{\max}(S) > t \,|\, \Delta(\mathrm{PA}(|T|, S)) = \Delta(T)) \le (1+o(1))c(|T|, \Delta(T))t^{1-2|T|+2\Delta(T)} \exp\bigl(-t^2/4\bigr)$$

as $t \to \infty$. Consequently we have that

$$\mathbb{P}(D_{\max}(S) > t) \le (1+o(1))\mathbb{P}(\Delta(\mathrm{PA}(|T|, S)) = \Delta(T)) \times c(|T|, \Delta(T))t^{1-2|T|+2\Delta(T)} \exp\bigl(-t^2/4\bigr)$$

as $t \to \infty$, which combined with the tail behavior of $D_{\max}(T)$ gives that

$$\mathbb{P}(D_{\max}(T) > t) - \mathbb{P}(D_{\max}(S) > t) \ge (1-o(1))\mathbb{P}(\Delta(\mathrm{PA}(|T|, S)) < \Delta(T)) \times c(|T|, \Delta(T))t^{1-2|T|+2\Delta(T)} \exp\bigl(-t^2/4\bigr)$$

as $t \to \infty$. To conclude the proof, notice that $\mathbb{P}(\Delta(\mathrm{PA}(|T|, S)) < \Delta(T))$ is at least as great as the probability that vertex $|S|+1$ connects to a leaf of S, which has probability at least $1/(2|S|-2)$.

Case 3: $|S| = |T|$, *different degree profiles.* Let $z \in \{1, \ldots, |T|\}$ be the first index such that $d_S(z) \neq d_T(z)$ and assume w.l.o.g. that $d_S(z) < d_T(z)$. First we have that

$$\mathbb{P}(D_{\max}(T) > t) \ge \mathbb{P}(\exists i \in [z-1] : D_i(T) > t) + \mathbb{P}(D_z(T) > t) - \sum_{i=1}^{z-1} \mathbb{P}(D_z(T) > t, D_i(T) > t)$$

and

$$\mathbb{P}(D_{\max}(S) > t) \le \mathbb{P}(\exists i \in [z-1] : D_i(S) > t) + \sum_{i=z}^{\infty} \mathbb{P}(D_i(S) > t).$$

Now observe that one can couple the evolution of $\mathrm{PA}(n, T)$ and $\mathrm{PA}(n, S)$ in such a way that the degrees of vertices $1, \ldots, z-1$ stay the same in both trees. Thus one clearly has

$$\mathbb{P}(\exists i \in [z-1] : D_i(T) > t) = \mathbb{P}(\exists i \in [z-1] : D_i(S) > t).$$

Putting the three above displays together one obtains

$$\mathbb{P}(D_{\max}(T) > t) - \mathbb{P}(D_{\max}(S) > t) \ge \mathbb{P}(D_z(T) > t) - \sum_{i=1}^{z-1} \mathbb{P}(D_z(T) > t, D_i(T) > t) - \sum_{i=z}^{\infty} \mathbb{P}(D_i(S) > t).$$

Now using Lemma 1 one easily gets (for some constant $C > 0$) that

$$\mathbb{P}(D_z(T) > t) \sim c(|T|, d_T(z))t^{1-2|T|+2d_T(z)} \exp\bigl(-t^2/4\bigr),$$

$$\sum_{i=1}^{z-1} \mathbb{P}(D_z(T) > t, D_i(T) > t) \le \sum_{i=1}^{z-1} \mathbb{P}(D_z(T) + D_i(T) > 2t) \le \sum_{i=1}^{z-1} (1+o(1))c(|T|, d_T(z) + d_T(i)) \times (2t)^{1-2|T|+2(d_T(z)+d_T(i))} \exp\bigl(-t^2\bigr),$$

$$\sum_{i=z}^{\infty} \mathbb{P}(D_i(S) > t) \le Ct^{1-2|T|+2d_S(z)} \exp\bigl(-t^2/4\bigr).$$

In particular, since $d_S(z) < d_T(z)$ and $t^\alpha \exp(-t^2) = o(\exp(-t^2/4))$ for any α, this shows that

$$\mathbb{P}(D_{\max}(T) > t) - \mathbb{P}(D_{\max}(S) > t) \ge (1-o(1))c(|T|, d_T(z))t^{1-2|T|+2d_T(z)} \exp\bigl(-t^2/4\bigr),$$

which, together with (12), concludes the proof. □

Proof of Theorem 3. As before we have that

$$\delta(S_k, T) \ge \sup_{t \ge 0} [\mathbb{P}(D_{\max}(S_k) > t) - \mathbb{P}(D_{\max}(T) > t)] \ge \mathbb{P}\Bigl(D_{\max}(S_k) > \sqrt{k}/2\Bigr) - \mathbb{P}\Bigl(D_{\max}(T) > \sqrt{k}/2\Bigr). \tag{13}$$

By Corollary 1, we know that the second term in (13) goes to zero as $k \to \infty$ for any fixed T. We can lower bound the first term in (13) by $\mathbb{P}\bigl(D_1(S_k) > \sqrt{k}/2\bigr) = 1 - \mathbb{P}\bigl(D_1(S_k) \le \sqrt{k}/2\bigr)$. From (4) we have that the first two moments of $D_1(S_k)$ are $\mathbb{E}D_1(S_k) = \Gamma(k)/\Gamma(k-1/2)$ and $\mathbb{E}D_1(S_k)^2 = \Gamma(k+1)/\Gamma(k) = k$. From standard facts about the Γ function and Stirling series one has that $0 \le \mathbb{E}D_1(S_k) - \sqrt{k-1} \le \bigl(6\sqrt{k-1}\bigr)^{-1}$ and then also

$$\mathbb{V}\mathrm{ar}\,(D_1(S_k)) = \mathbb{E}D_1(S_k)^2 - (\mathbb{E}D_1(S_k))^2 \le k - (k-1) = 1.$$

Therefore Chebyshev's inequality implies that $\lim_{k\to\infty} \mathbb{P}\bigl(D_1(S_k) \le \sqrt{k}/2\bigr) = 0$. □

3.2 Towards a Proof of Conjecture 1

Our proof of Theorem 2 above relied on the precise asymptotic tail behavior of $D_{\max}(T)$, as described in Corollary 1. In order to distinguish two trees with the same degree profile (such as the pair of trees in Fig. 1), it is necessary to incorporate information about the graph structure. Indeed, if S and T have the same degree profiles, then it is possible to couple $\mathrm{PA}(n, S)$ and $\mathrm{PA}(n, T)$ such that they have the same degree profiles for every n.

Thus a possible way to prove Conjecture 1 is to generalize the notion of maximum degree in a way that incorporates information about the graph structure, and then use similar arguments as in the proofs above. A candidate is the following.

Definition 1. *Given a tree U, define the* U-maximum degree *of a tree T, denoted by $\Delta_U(T)$, as*

$$\Delta_U(T) = \max_{\varphi} \sum_{u \in V(U)} d_T(\varphi(u)),$$

where $V(U)$ denotes the vertex set of U, and the maximum is taken over all injective graph homomorphisms from U to T. That is, φ ranges over all injective maps from $V(U)$ to $V(T)$ such that $\{u, v\} \in E(U)$ implies that $\{\varphi(u), \varphi(v)\} \in E(T)$, where $E(U)$ denotes the edge set of U, and $E(T)$ is defined similarly.

When U is a single vertex, then $\Delta_U \equiv \Delta$, so this indeed generalizes the notion of maximum degree.

Intuitively, the main contributor to the tail of $\Delta_T(\mathrm{PA}(n,T))$ should be the homomorphism that maps T to the vertices making up the initial seed. In other words, the tail should behave like the tail of the sum of the degrees of the initial vertices. On the other hand, if S is not isomorphic to T (and assume for simplicity that $|S| = |T|$), then any homomorphism from T to $\mathrm{PA}(n,S)$ must use a vertex that is not part of the seed. Because of this, one expects that the tail of $\Delta_T(\mathrm{PA}(n,S))$ is lighter than the tail of $\Delta_T(\mathrm{PA}(n,T))$. In particular, we conjecture the following.

Conjecture 2. Suppose S and T are two non-isomorphic trees of the same size. Then

$$\limsup_{n\to\infty} \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,S)) > t\sqrt{n}\big) = o\big(t^{2|T|-3}\exp\big(-t^2/4\big)\big)$$

as $t \to \infty$.

If this conjecture were true, then Conjecture 1 also follows, as we now show.

Proof of Conjecture 1 assuming Conjecture 2 holds. Assume $|S| = |T|$; if $|S| \neq |T|$ we already know from Theorem 2 that $\delta(S,T) > 0$. As in the proof of Theorem 2, for any $t > 0$ and $n \geq \max\{|S|, |T|\}$ we have that

$$\begin{aligned}
&\mathrm{TV}(\mathrm{PA}(n,S), \mathrm{PA}(n,T)) \\
&\quad \geq \mathrm{TV}(\Delta_T(\mathrm{PA}(n,S)), \Delta_T(\mathrm{PA}(n,T))) \\
&\quad \geq \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,T)) > t\sqrt{n}\big) - \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,S)) > t\sqrt{n}\big),
\end{aligned}$$

and consequently

$$\begin{aligned}
\delta(S,T) \geq \sup_{t>0} \Big\{ &\liminf_{n\to\infty} \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,T)) > t\sqrt{n}\big) \\
&- \limsup_{n\to\infty} \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,S)) > t\sqrt{n}\big)\Big\}.
\end{aligned} \tag{14}$$

Since $\varphi(i) = i$ for $1 \leq i \leq |T|$ is an injective graph homomorphism from T to $\mathrm{PA}(n,T)$, we have that

$$\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\big(\Delta_T(\mathrm{PA}(n,T)) > t\sqrt{n}\big) \\
&\quad \geq \liminf_{n\to\infty} \mathbb{P}\left(\sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i) > t\sqrt{n}\right) = \mathbb{P}\left(\sum_{i=1}^{|T|} D_i(T) > t\right).
\end{aligned}$$

By Lemma 1 we know that

$$\mathbb{P}\left(\sum_{i=1}^{|T|} D_i(T) > t\right) \sim c(|T|, 2|T|-2)t^{2|T|-3}\exp\big(-t^2/4\big)$$

as $t \to \infty$, which together with (14) and Conjecture 2 shows that $\delta(S,T) > 0$. □

We note that the statistics considered by [7] are very similar to the ones considered above based on the U-maximum degree. More precisely, instead of taking a maximum over homomorphisms, they take a sum over them, and instead of taking a sum over vertices, they take a product over them. (They also consider decorated trees, which essentially means raising the degrees appearing in the statistic to appropriate powers.) Furthermore, while we considered the tail behavior of statistics based on the U-maximum degree, they constructed appropriate martingales, for which they needed to estimate the first two moments of these statistics. Understanding the connection between these two related approaches would be interesting.

4 The Weak Limit of $\mathrm{PA}(n,T)$

In this section we prove Theorem 1. For two graphs G and H we write $G = H$ if G and H are isomorphic, and we use the same notation for rooted graphs. Recalling the definition of the Benjamini-Schramm limit (see [Definition 2.1., [4]]), we want to prove that

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(B_r(\mathrm{PA}(n,T), k_n(T)) = (H,y)) \\
&\quad = \mathbb{P}(B_r(\mathcal{T},(0)) = (H,y)),
\end{aligned}$$

where $B_r(G,v)$ is the rooted ball of radius r around vertex v in the graph G, $k_n(T)$ is a uniformly random vertex in $\mathrm{PA}(n,T)$, (H,y) is a finite rooted tree and $(\mathcal{T},(0))$ is the Pólya-point graph (with $m = 1$).

We construct a forest F based on T as follows. To each vertex v in T we associate $d_T(v)$ isolated nodes with self loops, that is F consists of $2(|T|-1)$ isolated vertices with self loops. Our convention here is that a node with k regular edges and one self loop has degree $k+1$. The graph evolution process $\mathrm{PA}(n,F)$ for forests is defined in the same way as for trees, and we couple the processes $\mathrm{PA}(n,T)$ and $\mathrm{PA}(n+|T|-2,F)$ in the natural way: when an edge is added to vertex v of T in $\mathrm{PA}(n,T)$ then an edge is also added to one of the $d_T(v)$ corresponding vertices of F in $\mathrm{PA}(n+|T|-2,F)$, and furthermore newly added vertices are always coupled. We first observe that, clearly, the weak limit of $\mathrm{PA}(n+|T|-2,F)$ is the Pólya-point graph, that is

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(B_r(\mathrm{PA}(n+|T|-2,F), k_n(F)) = (H,y)) \\
&\quad = \mathbb{P}(B_r(\mathcal{T},(0)) = (H,y)),
\end{aligned}$$

where $k_n(F)$ is a uniformly random vertex in $\mathrm{PA}(n+|T|-2,F)$. We couple $k_n(F)$ and $k_n(T)$ in the natural way, that is if $k_n(F)$ is the tth newly created vertex in $\mathrm{PA}(n+|T|-2,F)$ then $k_n(T)$ is the tth newly created vertex in $\mathrm{PA}(n,T)$. To conclude the proof it is now sufficient to show that

$$\lim_{n\to\infty} \mathbb{P}\big(B_r(\mathrm{PA}(n+|T|-2,F),k_n(F)) \neq B_r(\mathrm{PA}(n,T),k_n(T))\big) = 0.$$

The following inequalities hold true (with a slight—but clear—abuse of notation when we write $v \in F$) for any $u > 0$,

$$\begin{aligned}
&\mathbb{P}\big(B_r(\mathrm{PA}(n+|T|-2,F),k_n(F)) \neq B_r(\mathrm{PA}(n,T),k_n(T))\big)\\
&\quad\le \mathbb{P}\big(\exists v \in F \text{ s.t. } v \in B_r(\mathrm{PA}(n+|T|-2,F),k_n(F))\big)\\
&\quad\le \mathbb{P}\big(\exists v \in F, d_{\mathrm{PA}(n+|T|-2,F)}(v) < u\big)\\
&\quad+ \mathbb{P}\big(\exists v \in B_r(\mathrm{PA}(n+|T|-2,F),k_n(F))\\
&\qquad\text{s.t.}\quad d_{\mathrm{PA}(n+|T|-2,F)}(v) \ge u\big).
\end{aligned}$$

It is easy to verify that for any $u > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\exists v \in F, d_{\mathrm{PA}(n+|T|-2,F)}(v) < u\big) = 0.$$

Furthermore since $B_r(\mathrm{PA}(n+|T|-2,F),k_n(F))$ tends to the Pólya-point graph we also have

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\big(\exists v \in B_r(\mathrm{PA}(n+|T|-2,F),k_n(F)) \text{ s.t. } d_{\mathrm{PA}(n+|T|-2,F)}(v) \ge u\big)\\
&\quad= \mathbb{P}\big(\exists v \in B_r(\mathcal{T},(0)) \text{ s.t. } d_{\mathcal{T}}(v) \ge u\big).
\end{aligned}$$

By looking at the definition of $(\mathcal{T},(0))$ given in [4] one can easily show that

$$\lim_{u\to\infty} \mathbb{P}\big(\exists v \in B_r(\mathcal{T},(0)) \text{ s.t. } d_{\mathcal{T}}(v) \ge u\big) = 0,$$

which concludes the proof.

5 Proof of Lemma 1

In this section we prove Lemma 1. In light of the representation (5) in Section 2.1.2, part (a) of Lemma 1 follows from a lengthy computation, the result of which we state separately.

Lemma 2. *Fix positive integers a and b. Let B and Z be independent random variables such that $B \sim \mathrm{Beta}(a,b)$ and $Z \sim \mathrm{GGa}(a+b+1,2)$, and let $V = 2BZ$. Then*

$$\mathbb{P}(V > t) \sim c\left(\frac{a+b+2}{2}, a\right) t^{-1+a-b} \exp\big(-t^2/4\big) \qquad (15)$$

as $t \to \infty$, where the constant c is as in (1).

Proof. By definition we have for $t > 0$ that

$$\begin{aligned}
\mathbb{P}(V > t) &= \mathbb{P}(2BZ > t)\\
&= \int_{t/2}^{\infty} \int_{t/(2z)}^{1} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx\\
&\qquad\times \frac{2}{\Gamma\left(\frac{a+b+1}{2}\right)} z^{a+b} e^{-z^2} dz\\
&= \int_{t/2}^{\infty} \big[1 - I_{t/(2z)}(a,b)\big] \frac{2}{\Gamma\left(\frac{a+b+1}{2}\right)} z^{a+b} e^{-z^2} dz,
\end{aligned}$$

where $I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x y^{a-1}(1-y)^{b-1} dy$ is the regularized incomplete Beta function. For positive integers a and b, integration by parts and induction gives that

$$I_x(a,b) = 1 - \sum_{j=0}^{a-1} \binom{a+b-1}{j} x^j (1-x)^{a+b-1-j}.$$

Plugging this back in to the integral and doing a change of variables $y = 2z$, we get that

$$\begin{aligned}
\mathbb{P}(V > t) = \frac{2^{-(a+b)}}{\Gamma\left(\frac{a+b+1}{2}\right)} &\sum_{j=0}^{a-1} \binom{a+b-1}{j}\\
&\times \int_t^{\infty} t^j (y-t)^{a+b-1-j} y \, \exp\big(-y^2/4\big) dy.
\end{aligned}$$

Expanding $(y-t)^{a+b-1-j}$ we arrive at the alternating sum formula

$$\begin{aligned}
\mathbb{P}(V > t) = \frac{2^{-(a+b)}}{\Gamma\left(\frac{a+b+1}{2}\right)} &\sum_{j=0}^{a-1} \sum_{k=0}^{a+b-1-j} \binom{a+b-1}{j} \binom{a+b-1-j}{k}\\
&\times (-1)^{a+b-1-j-k} t^{a+b-1-k} A_{k+1},
\end{aligned} \qquad (16)$$

where for $m \ge 0$ let

$$A_m := \int_t^{\infty} y^m \exp\big(-y^2/4\big) dy.$$

Thus in order to show (15) it is enough to show that for every j such that $0 \le j \le a-1$ we have

$$\begin{aligned}
&\sum_{k=0}^{a+b-1-j} \binom{a+b-1-j}{k} (-1)^{a+b-1-j-k} t^{a+b-1-k} A_{k+1}\\
&\quad\sim \frac{2^{a+b-j}(a+b-1-j)!}{t^{a+b-1-2j}} \exp\big(-t^2/4\big).
\end{aligned} \qquad (17)$$

To do this, we need to evaluate the integrals $\{A_m\}_{m\ge 0}$. Recall that the complementary error function is defined as $\mathrm{erfc}(z) = 1 - \mathrm{erf}(z) = (2/\sqrt{\pi}) \int_z^{\infty} \exp(-u^2) du$, and thus $A_0 = \sqrt{\pi}\,\mathrm{erfc}(t/2)$; also $A_1 = 2\,\exp(-t^2/4)$. Integration by parts gives that for $m \ge 2$ we have $A_m = 2t^{m-1}\,\exp(-t^2/4) + 2(m-1)A_{m-2}$. Iterating this, and using the values for A_0 and A_1, gives us that for m odd we have

$$A_m = 2t^{m-1} \exp\big(-t^2/4\big) \sum_{\ell=0}^{\frac{m-1}{2}} \frac{(m-1)!!}{(m-2\ell-1)!!} \left(\frac{2}{t^2}\right)^{\ell}, \qquad (18)$$

and for m even we have

$$\begin{aligned}
A_m = 2t^{m-1} \exp\big(-t^2/4\big) &\sum_{\ell=0}^{\frac{m}{2}-1} \frac{(m-1)!!}{(m-2\ell-1)!!} \left(\frac{2}{t^2}\right)^{\ell}\\
&+ 2^{\frac{m}{2}} \times (m-1)!! \times \sqrt{\pi}\,\mathrm{erfc}(t/2).
\end{aligned} \qquad (19)$$

In the following we fix j such that $0 \le j \le a-1$ and $a+b-1-j$ is odd—showing (17) when $a+b-1-j$ is even can be done in the same way. In order to abbreviate notation we let $r = (a+b-2-j)/2$. Plugging in the formulas (18) and (19) into the left hand side of (17) we get that

$$\begin{aligned}
&\sum_{k=0}^{a+b-1-j}\binom{a+b-1-j}{k}(-1)^{a+b-1-j-k}t^{a+b-1-k}A_{k+1}\\
&\quad=\sum_{k=0}^{2r+1}\binom{2r+1}{k}(-1)^{2r+1-k}t^{2r+1+j-k}A_{k+1}\\
&\quad=-\sum_{\ell=0}^{r}\binom{2r+1}{2\ell}t^{2r+1+j-2\ell}A_{2\ell+1}\\
&\qquad+\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}t^{2r+1+j-(2\ell+1)}A_{2\ell+2}\\
&\quad=-\sum_{\ell=0}^{r}\binom{2r+1}{2\ell}t^{2r+1+j-2\ell}2\exp\bigl(-t^2/4\bigr)\\
&\qquad\times\sum_{u=0}^{\ell}2^u\frac{(2\ell)!!}{(2\ell-2u)!!}t^{2\ell-2u}\\
&\qquad+\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}t^{2r+1+j-(2\ell+1)}2\exp\bigl(-t^2/4\bigr)\\
&\qquad\times\sum_{u=0}^{\ell}2^u\frac{(2\ell+1)!!}{(2\ell+1-2u)!!}t^{2\ell+1-2u}\\
&\qquad+\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}t^{2r+1+j-(2\ell+1)}2^{\ell+1}(2\ell+1)!!\sqrt{\pi}\operatorname{erfc}(t/2)\\
&\quad=2\exp\bigl(-t^2/4\bigr)\sum_{u=0}^{r}t^{2r+1+j-2u}2^u\\
&\qquad\times\sum_{k=2u}^{2r+1}\binom{2r+1}{k}(-1)^{k+1}\frac{k!!}{(k-2u)!!}
\end{aligned}\tag{20}$$

$$+\sqrt{\pi}\operatorname{erfc}(t/2)\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}t^{2r+1+j-(2\ell+1)}2^{\ell+1}(2\ell+1)!!.\tag{21}$$

An important fact that we will use is that for every polynomial P with degree less than n we have

$$\sum_{k=0}^{n}\binom{n}{k}(-1)^kP(k)=0.\tag{22}$$

Consequently, applying this to the polynomial $P(k)=k(k-2)\cdots(k-2(u-1))$ we get that

$$\begin{aligned}
&\sum_{k=2u}^{2r+1}\binom{2r+1}{k}(-1)^{k+1}k(k-2)\cdots(k-2(u-1))\\
&\quad=\sum_{k=0}^{2u-1}\binom{2r+1}{k}(-1)^{k}k(k-2)\cdots(k-2(u-1))\\
&\quad=-\sum_{\ell=0}^{u-1}\binom{2r+1}{2\ell+1}(2\ell+1)(2\ell-1)\cdots(2\ell+1-2(u-1))\\
&\quad=-\sum_{\ell=0}^{u-1}\binom{2r+1}{2\ell+1}(2\ell+1)!!(2(u-1-\ell)-1)!!(-1)^{u-1-\ell}.
\end{aligned}\tag{23}$$

Thus we see that in the sum (20) the cofficient of the term involving t^{2r+1+j} is zero, while the coefficient of the term involving $t^{2r+1+j-2u}$ for $1\le u\le r$ is $2^{u+1}\exp(-t^2/4)$ times the expression in (23). These are cancelled by terms coming from the sum in (21) as we will see shortly; to see this we need the asymptotic expansion of erfc to high enough order. In particular we have (see [1, Equations 7.1.13 and 7.1.24]) that

$$\begin{aligned}
\sqrt{\pi}\operatorname{erfc}(t/2)&=2\exp\bigl(-t^2/4\bigr)\\
&\quad\times\sum_{n=0}^{2r}(-1)^n2^n(2n-1)!!t^{-2n-1}+R(t),
\end{aligned}\tag{24}$$

where the approximation error $R(t)$ satisfies

$$|R(t)|\le 2^{2r+2}(4r+1)!!t^{-(4r+3)}\exp\bigl(-t^2/4\bigr).$$

Plugging (24) back into (21), we first see that the error term satisfies

$$\begin{aligned}
&|R(t)|\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}t^{2r+1+j-(2\ell+1)}2^{\ell+1}(2\ell+1)!!\\
&\quad=O(t^{2j-1-(a+b)}\exp(-t^2/4))
\end{aligned}\tag{25}$$

as $t\to\infty$. The main term of (21) becomes the sum

$$\begin{aligned}
&2\exp\bigl(-t^2/4\bigr)\sum_{\ell=0}^{r}\sum_{n=0}^{2r}\binom{2r+1}{2\ell+1}2^{\ell+n+1}\\
&\qquad\times(2\ell+1)!!(2n-1)!!(-1)^nt^{2r+1+j-2(\ell+n+1)}.
\end{aligned}$$

For u such that $1\le u\le r$, the coefficient of the term involving $t^{2r+1+j-2u}$ is $2^{u+1}\exp(-t^2/4)$ times

$$\sum_{\ell=0}^{u-1}\binom{2r+1}{2\ell+1}(2\ell+1)!!(2(u-1-\ell)-1)!!(-1)^{u-1-\ell},$$

which cancels out the coefficient of the same term coming from the other sum (20), see (23). For u such that $r<u\le 2r$, the coefficient of the term involving $t^{2r+1+j-2u}$ is $2^{u+1}\exp(-t^2/4)$ times

$$\begin{aligned}
&\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}(2\ell+1)!!(2(u-1-\ell)-1)!!(-1)^{u-1-\ell}\\
&\quad=\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}(2\ell+1)(2\ell-1)\ldots((2\ell+1)-2(u-1))\\
&\quad=-\sum_{k=0}^{2r+1}\binom{2r+1}{k}(-1)^kk(k-2)\ldots(k-2(u-1))=0,
\end{aligned}$$

where we again used (22), together with the fact that $u\le 2r$. Finally, the coefficient of the term involving $t^{2j+1-(a+b)}$ is $2^{2r+2}\exp(-t^2/4)$ times

$$\begin{aligned}
&\sum_{\ell=0}^{r}\binom{2r+1}{2\ell+1}(2\ell+1)!!(2(2r-\ell)-1)!!(-1)^{2r-\ell}\\
&\quad=-\sum_{k=0}^{2r+1}\binom{2r+1}{k}(-1)^kk(k-2)\ldots(k-4r)\\
&\quad=-\sum_{k=0}^{2r+1}\binom{2r+1}{k}(-1)^kk^{2r+1}=-(-1)^{2r+1}(2r+1)!\\
&\quad=(2r+1)!,
\end{aligned}$$

where we used (22) in the second equality. Since all other terms are of lower order (see (25)), this concludes the proof. □

Proof of Lemma 1. (a) If $U \neq T$, then $d = \sum_{i\in U} d_T(i) \in \{1, \ldots, 2|T| - 3\}$. Similarly to the third paragraph in Section 2.1.2, we can view the evolution of $\sum_{i\in U} d_{\mathrm{PA}(n,T)}(i)$ in the following way. When adding a new vertex, first decide whether it attaches to one of the initial $|T|$ vertices (with probability $\sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i)/(2n-2)$) or not (with the remaining probability); if it does, then independently pick one of them to attach to with probability proportional to their degree—a vertex in U is chosen with probability $\sum_{i\in U} d_{\mathrm{PA}(n,T)}(i)/\sum_{i=1}^{|T|} d_{\mathrm{PA}(n,T)}(i)$. This implies the following representation: $\sum_{i\in U} D_i(T) \stackrel{d}{=} 2BZ$, where B and Z are independent, $B \sim \mathrm{Beta}(d, 2|T| - 2 - d)$, and $Z \sim \mathrm{GGa}(2|T|- 1, 2)$. This also follows directly from the representation (5). Thus (6) is a direct consequence of Lemma 2.

If $U = T$, then $\sum_{i\in U} D_i(T) \stackrel{d}{=} 2Z$ where $Z \sim \mathrm{GGa}(2|T| - 1, 2)$ (see Section 2.1.2), and then (6) follows from a calculation that is contained in the proof of Lemma 2.

(b) To show (7) we use the results of [13] as described in Section 2.1.1. In addition we use the following tail bound of [13, Lemma 2.6], which says that for $x > 0$ and $s \geq 1$ we have $\int_x^\infty \kappa_s(y)dy \leq \frac{s}{x}\kappa_s(x)$. Consequently, for any $i > |T|$ we have the following tail bound:

$$\begin{aligned}\mathbb{P}(D_i(T) > t) &= \mathbb{P}\left(W_{i-1} > \sqrt{\frac{i-1}{2}}t\right) = \int_{\sqrt{\frac{i-1}{2}}t}^\infty \kappa_{i-1}(y)dy \\ &\leq \frac{\sqrt{2i-2}}{t}\kappa_{i-1}\left(\sqrt{\frac{i-1}{2}}t\right) \\ &= \frac{2}{\sqrt{\pi}t}\,\exp\bigl(-t^2/4\bigr)(i-2)!U\left(i-2,\frac{1}{2},\frac{t^2}{4}\right).\end{aligned}$$

The following integral representation is useful for us [1, eq. 13.2.5]:

$$\Gamma(a)U(a,b,z) = \int_0^\infty e^{-zw}w^{a-1}(1+w)^{b-a-1}dw.$$

Consequently, we have

$$\begin{aligned}&\sum_{i=3}^\infty (i-2)!U\left(i-2,\frac{1}{2},\frac{t^2}{4}\right) \\ &\quad= \int_0^\infty e^{-\frac{t^2}{4}w}\frac{1}{w\sqrt{1+w}}\sum_{i=3}^\infty (i-2)\left(\frac{w}{1+w}\right)^{i-2}dw \\ &\quad= \int_0^\infty e^{-\frac{t^2}{4}w}\frac{1}{w\sqrt{1+w}}w(1+w)dw \\ &\quad\leq \int_0^\infty e^{-\frac{t^2}{4}w}(1+w)dw = \frac{4}{t^2}+\frac{16}{t^4},\end{aligned}$$

which shows (7) for $L = 3$. Similarly, for $L \geq 4$ we have

$$\begin{aligned}&\sum_{i=L}^\infty (i-2)!U\left(i-2,\frac{1}{2},\frac{t^2}{4}\right) \\ &\quad= \int_0^\infty e^{-\frac{t^2}{4}w}\frac{1}{w\sqrt{1+w}}\sum_{i=L}^\infty (i-2)\left(\frac{w}{1+w}\right)^{i-2}dw \\ &\quad= \int_0^\infty e^{-\frac{t^2}{4}w}\frac{1}{w\sqrt{1+w}}\,\frac{(L-2)\left(\frac{w}{1+w}\right)^{L-2}+(3-L)\left(\frac{w}{1+w}\right)^{L-1}}{1/(1+w)^2}dw \\ &\quad\leq \int_0^\infty e^{-\frac{t^2}{4}w}(L-2)\left(\frac{w}{1+w}\right)^{L-3}\sqrt{1+w}dw \\ &\quad\leq \int_0^\infty e^{-\frac{t^2}{4}w}(L-2)w^{L-3}dw = \frac{4^{L-2}\times(L-2)!}{t^{2L-4}},\end{aligned}$$

where the first inequality follows from dropping the nonpositive term $(3-L)(\frac{w}{1+w})^{L-1}$, and the second one follows because $L \geq 4$. This shows (7) for $L \geq 4$ and thus concludes the proof. □

6 Open Problems

1) This paper is essentially about the *testing* version of the problem. Can anything be said about the *estimation* version? Perhaps a first step would be to understand the multiple hypothesis testing problem where one is interested in testing whether the seed belongs to the family of trees $\mathcal{T}_1$ or to the family $\mathcal{T}_2$.
2) Starting from two seeds S and T with different spectrum, is it always possible to distinguish (with nontrivial probability) between $\mathrm{PA}(n, S)$ and $\mathrm{PA}(n, T)$ with spectral techniques? More generally, it would be interesting to understand what properties are invariant under modifications of the seed.
3) Is it possible to give a combinatorial description of the metric δ?
4) Under what conditions on two tree sequences (T_k), (R_k) do we have $\lim_{k\to\infty}\delta(T_k, R_k) = 1$? In Theorem 3 we showed that a sufficient condition is to have $T_k = T$ and $R_k = S_k$. This can easily be extended to the condition that $\Delta(T_k)$ remains bounded while $\Delta(R_k)$ tends to infinity. If T_k and R_k are independent (uniformly) random trees on k vertices, do we have $\lim_{k\to\infty}\mathbb{E}\delta(T_k, R_k) = 1$?
5) What can be said about the general preferential attachment model, when multiple edges or vertices are added at each step?
6) A simple variant on the model studied in this paper is to consider probabilities of connection proportional to the degree of the vertex raised to some power α. For $\alpha = 1$ our results and those of [7] show that different seeds are distinguishable. What about for other α?

   In forthcoming work [6], we show that for $\alpha = 0$, i.e., in the case of uniform attachment, the same result holds: different seeds are distinguishable, in the sense that $\delta_0(S, T) > 0$ when S and T are nonisomorphic trees on at least three vertices (here $\delta_\alpha(S, T)$ is defined analogously to $\delta(S, T)$ for general α). What about $\alpha \in (0, 1)$, is $\delta_\alpha(S, T) > 0$? What can be said

about $\delta_\alpha(S,T)$ as a function of α? Is it monotone in α? Is it convex?

When $\alpha > 1$, i.e., in the case of superlinear preferential attachment, we expect the seed to have an influence in the strongest sense, i.e., that if S and T are nonisomorphic trees on at least three vertices, then

$$\mathrm{TV}(\mathrm{PA}_\alpha(\infty, S), \mathrm{PA}_\alpha(\infty, T)) > 0. \tag{26}$$

When $\alpha > 1$, [12] give a precise description of the infinite tree $\mathrm{PA}_\alpha(\infty, S_2)$, which contains exactly one vertex of infinite degree, with all other vertices having finite degree. From this it is possible to give a similar description of the infinite tree $\mathrm{PA}_\alpha(\infty, S)$ for any seed tree S. We believe that from this description it is possible to deduce that (26) holds for $\alpha > 1$, but have not pursued this question further.

Acknowledgments

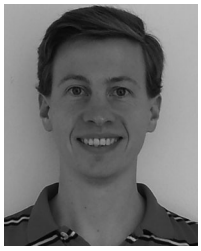
The first author thanks Nati Linial for initial discussions on this problem. They also thank Remco van der Hofstad and Nathan Ross for helpful discussions and valuable pointers to the literature. The research described here was carried out at the Simons Institute for the Theory of Computing. We are grateful to the Simons Institute for offering us such a wonderful research environment. This work was supported by NSF grants DMS 1106999 and CCF 1320105, by grant 328025 from the Simons Foundation (E.M) and by DOD ONR grants N000141110140 and N00014-14-1-0823 (E.M., M.Z.R.). Miklós Z. Rácz is the corresponding author.

References

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, vol. 55. New York, NY, USA: Dover, 1964.
[2] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, 1999.
[3] I. Benjamini and O. Schramm, "Recurrence of distributional limits of finite planar graphs," *Electron. J. Probability*, vol. 6, no. 23, pp. 1–13, 2001.
[4] N. Berger, C. Borgs, J. T. Chayes, and A. Saberi, "Asymptotic behavior and distributional limits of preferential attachment graphs," *Ann. Probability*, vol. 42, no. 1, pp. 1–40, 2014.
[5] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády, "The degree sequence of a scale-free random graph process," *Random Structures Algorithms*, vol. 18, no. 3, pp. 279–290, 2001.
[6] S. Bubeck, R. Eldan, E. Mossel, and M. Z. Rácz, "From trees to seeds: On the inference of the seed from large trees in the uniform attachment model," *arXiv preprint arXiv:1409.7685*, 2014.
[7] N. Curien, T. Duquesne, I. Kortchemski, and I. Manolescu, "Scaling limits and influence of the seed graph in preferential attachment trees," *arXiv preprint arXiv:1406.1758*, 2014.
[8] S. Janson, "Limit theorems for triangular urn schemes," *Probability Theory Related Fields*, vol. 134, no. 3, pp. 417–452, 2006.
[9] R. D. Kleinberg and J. M. Kleinberg, "Isomorphism and embedding problems for infinite limits of scale-free graphs," in *Proc. 16th Annu. ACM-SIAM Symp. Discrete Algorithms*, 2005, pp. 277–286.
[10] H. M. Mahmoud, "Distances in random plane-oriented recursive trees," *J. Comput. Appl. Math.*, vol. 41, nos. 1/2, pp. 237–245, 1992.
[11] T. F. Móri, "The maximum degree of the Barabási-Albert random tree," *Combinatorics, Probability Comput.*, vol. 14, no. 03, pp. 339–348, 2005.
[12] R. Oliveira and J. Spencer, "Connectivity transitions in networks with super-linear preferential attachment," *Internet Math.*, vol. 2, no. 2, pp. 121–163, 2005.
[13] E. A. Peköz, A. Röllin, and N. Ross, "Degree asymptotics with rates for preferential attachment random graphs," *Ann. Appl. Probability*, vol. 23, no. 3, pp. 1188–1218, 2013.
[14] E. A. Peköz, A. Röllin, and N. Ross, "Joint degree distributions of preferential attachment random graphs," *arXiv preprint arXiv:1402.4686*, 2014.
[15] A. Rudas, B. Tóth, and B. Valkó, "Random trees and general branching processes," *Random Structures Algorithms*, vol. 31, no. 2, pp. 186–202, 2007.
[16] R. Schweiger, M. Linial, and N. Linial, "Generative probabilistic models for protein-protein interaction networks—the biclique perspective," *Bioinformatics*, vol. 27, no. 13, pp. i142–i148, 2011.

Sebastien Bubeck is an assistant professor at ORFE, Princeton University, and a researcher in the Theory Group at Microsoft Research, Redmond.

Elchanan Mossel is a professor of statistics and computer science at University of Pennsylvania and U.C. Berkeley. He is interested in combinatorial statistics, discrete Fourier analysis and influences, randomized algorithms, computational complexity, MCMC, Markov random fields, social choice, game theory and evolution. He is Miki's advisor. Sebastien, Miki and Elchanan started to collaborate at the Simons Institute for the Theory of Computing at U.C. Berkeley.



Miklos Racz is a PhD student in statistics at UC Berkeley, advised by Elchanan Mossel. He graduated from the Budapest University of Technology and Economics in 2010 with an MS in mathematics, and he earned an MS in computer science from Berkeley in 2014. Miki is interested in probability theory and its applications, including topics such as networks, voting, interacting particle systems, game theory, and mathematical biology.