Statistical Risk Bounds for Genealogical Reconstruction on Random Recursive Trees

Vincent Counathe
Supervised by Christophe Giraud

Université Paris-Saclay, Institut de Mathématiques d'Orsay

We derive non-asymptotic upper and lower bounds for a statistical risk functional in genealogical reconstruction on random recursive trees. The analysis combines analytic combinatorics and martingale concentration techniques, with applications to inference in networked data.

August 2024

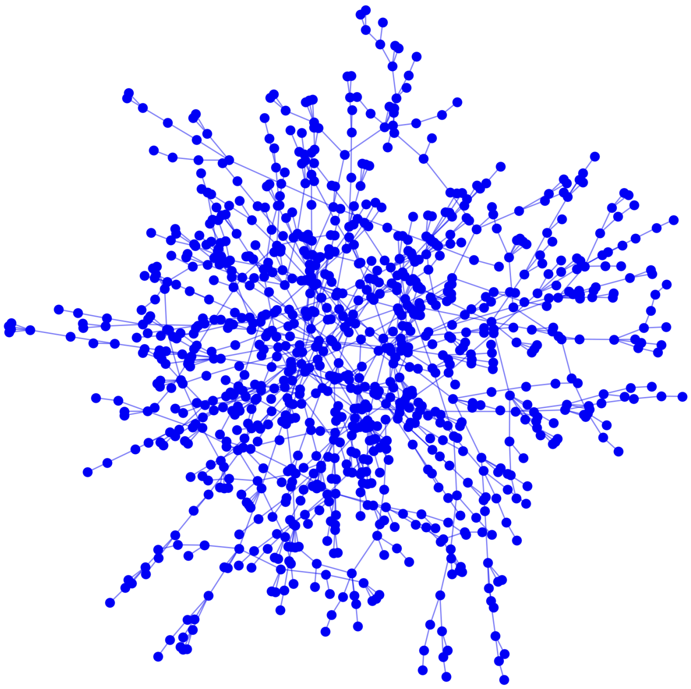

Figure 1: Uniform attachment model.

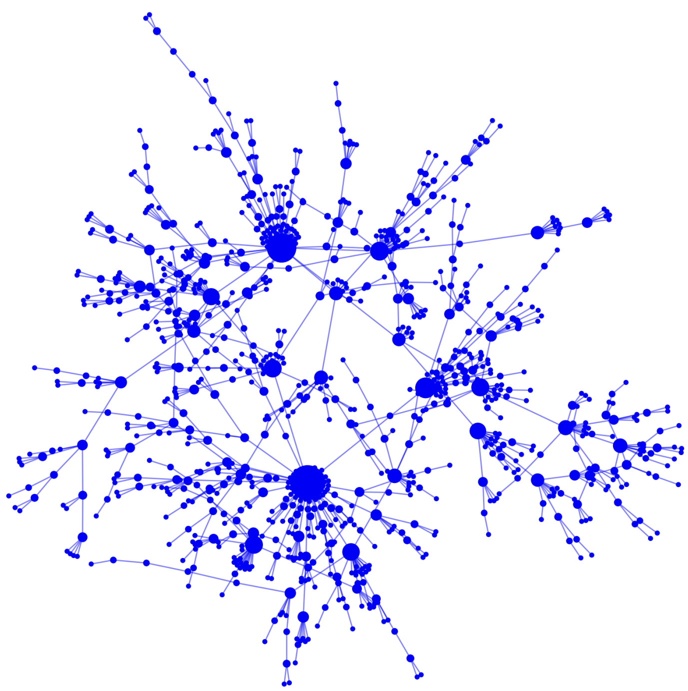

Figure 2: Preferential attachment model.

Acknowledgments

I am deeply grateful to Christophe Giraud for suggesting this fascinating topic and for his patient and insightful guidance throughout the course of this research project.

Contents

1 Introduction

1.1 Overview and Motivation

Random trees are a fundamental object of study in probability theory, combinatorics, and various applied fields such as computer science or biology. A random tree is a tree—a connected acyclic graph—whose structure is determined by some random procedure. These trees can serve as models for a wide range of natural and artificial phenomena: the spread of information, the growth of networks, or the evolutionary history of species.

One of the simplest models of a random tree is the uniform random recursive tree (URRT), where nodes are added one at a time, with each new node connecting to an existing node chosen uniformly at random. Other models commonly studied include the Galton-Watson process (see for instance the work of Neveu [9], Abraham and Delmas [1]) and preferential attachment model (PA), whereby new nodes connect with a higher probability to high-degree nodes, effectively mirroring the "rich-get-richer" phenomenon (see Barabási and Albert [3]). These various models help provide insight into the typical properties and substructures of large networks.

Random trees can also be studied through the combinatorial angle. The field of analytic combinatorics (see Flajolet and Sedgewick [7] or Drmota [6]) offers powerful tools to study the behavior of these trees. While these aspects are fascinating, they lie slightly outside the primary focus of this work. As such, only a few basic properties are reminded here, serving the proof of a short lemma.

Recent work has also focused on scaling limits of discrete random trees (see for instance the continuum random tree introduced by Aldous [2]), albeit this is not a topic of focus here.

Estimating the history of a random tree finds various applications in biology or network theory. The motivation is the following: given only the final structure of the tree, how can we reconstruct the process by which the tree was formed? This cannot be done in a deterministic way (indeed, vertices 1 and 2 are indistinguishable in a tree of size 2) and requires probability tools.

This work builds on the recent paper from Briend et al. [4], who explored the estimation of a tree's history by analyzing upper and lower bounds associated with a given risk measure and ordering procedure.

1.2 Notations

The URRT model describes a growth process where each new node is added by attaching it to any existing node in the tree, chosen uniformly at random. Specifically, starting with a single root node, each subsequent node i (where $i \geq 2$) connects to a node chosen uniformly at random.

The PA model describes a growth process where each new node "prefers" to attach to existing nodes with higher degrees. Specifically, the probability that a new node connects to an existing node i is proportional to the degree d_i of node i. The degree of a vertex is its number of neighbours (or connections).

The Affine Attachment (AA) model generalizes the URRT and PA model by introducing a linear term: the probability that a new node attaches to an existing node i is proportional to $d_i + a$, where $a \geq 0$. When $a = 0$, one obtains the PA model. When $a \to \infty$, one obtains the URRT model.

Vertices of a tree may be assigned two distinct labels: an arbitrary label, which has no inherent intuitive significance, and a second label that denotes its rank or time of arrival. The arbitrary label serves to identify vertices within a tree where the historical sequence is unknown. The problem of reconstructing the tree's history then amounts to find the unique permutation σ that assigns to each vertex's arbitrary label its time of arrival.

An ordering estimator $\hat{\sigma}$ is a permutation that estimates a time of arrival for each vertex, based on observed data. $\hat{\sigma}(i)$ hence denotes the rank of a vertex. The goal of an ordering estimator is to approximate the original order in which nodes were added to the tree.

A label-invariant estimator is an estimator which only depends on the structure of the tree, and is independent of the arbitrary labelling. Formally, it verifies, for any tree T and (unknown) permutation σ,

$$\hat{\sigma}(T, \sigma) \stackrel{d}{=} \hat{\sigma}(T^{\sigma'}, \sigma \circ (\sigma')^{-1}),$$

for a given permutation σ'. $T^{\sigma'}$ denotes the tree with label i replaced by $\sigma'(i)$.

Given a vertex j in $[n] := [1, n]$, a real parameter $\alpha > 0$, and an estimator $\hat{\sigma}$, we define

$$R_{\alpha,j}(\hat{\sigma}) := \mathbb{E}\left[\frac{|\hat{\sigma}(j) - \sigma(j)|}{\sigma(j)^{\alpha}}\right].$$

We will mostly look at the case $\alpha = 1$ in the following pages, i.e., the estimation error of a vertex's rank is normalized by said rank. Intuitively, this reflects the fact that arrival times of leaves are harder to estimate.

We use the notation $f(n) \lesssim g(n)$ to indicate that there exists a constant C such that $f(n) \leq C \cdot g(n)$ for sufficiently large n, and similarly $f(n) \gtrsim g(n)$ to denote that there exists a constant C such that $f(n) \geq C \cdot g(n)$ for sufficiently large n.

1.3 Ordering Procedures

We define here the two main ordering procedures in use in this paper. Definitions are in line with those of Briend et al. [4]. The first ordering we will look into is based on the Jordan centrality. It measures the centrality of a vertex u in a tree T based on the size of its subtrees. For a vertex u, the Jordan centrality is defined as:

$$\psi_T(u) = \max_{v \in T, v \sim u} |(T, u)_v|$$

where $(T, u)_v$ represents the size of the subtree rooted at v after removing u, and $v \sim u$ indicates that v is adjacent to u. Intuitively, a vertex with a small Jordan centrality tends to be located at the center of the tree. On the contrary, leaves (i.e., vertices with degree equal to one) have the largest Jordan centralities.

A centroid of a tree T is a vertex that minimizes the maximum size of any of its subtrees. Formally, a vertex c is a centroid if $c = \operatorname{argmin}_{u \in T} \psi_T(u)$. Any tree has at least one and at most two centroids.

The Jordan ordering sorts vertices of a tree by increasing values of ψ, with ties resolved randomly. We note $\hat{\sigma}_J$ the resulting estimator, associated with the Jordan ordering of a tree T_n. This estimator is label-invariant.

The second ordering which will be of interest is the descendant ordering used by Briend et al. [4]. Given a tree T_n and a vertex u, the descendant centrality is defined as follows:

$$\psi'_T(u) = n - \mathrm{de}(u),$$

where $\mathrm{de}(u)$ denotes the number of descendants of u. If the centroid c and vertex 1 (the root) coincide, then the descendant ordering and the Jordan ordering coincide (assuming ties are broken in the same deterministic way). When the root is unknown, which is the case here, this ordering cannot be practically implemented.

Sorting vertices by increasing values of ψ'_T results in an ordering denoted as $\hat{\sigma}'$.

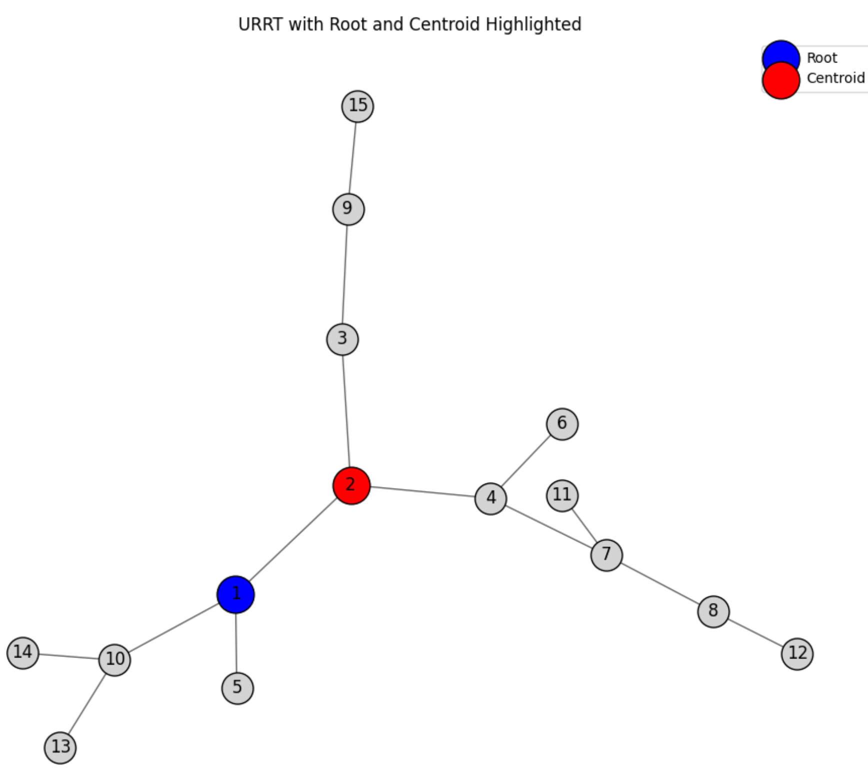

Figure 3: An illustration of a random tree following the URRT model, with $n = 15$.

We end the introduction with a lemma which will prove useful later on.

Lemma 1.1 (From Briend et al. [4]). *Given a tree T, let $c \in [n]$ denote a centroid of T, and let $\{1 \to c\}$ be the set of vertices on the path connecting 1 to c in T. Then for any $v \in [n] \setminus \{1 \to c\}$, we have:*

$$\psi_T(v) = \psi'_T(v).$$

2 Combinatorics Preliminaries

We introduce this brief section on combinatorics to present a lemma which will prove useful later in this paper. Given a tree T_n of size n, i.e., with n vertices, let us define $D_n := d(1, c)$, i.e., the distance between the root and the centroid in T_n. D_n is a random variable, and our aim is to upper bound $\mathbb{E}[D_n]$.

Let $\mathcal{T}$ denote the set of finite rooted trees, and $\mathcal{T}_n$ the set of finite rooted trees of size n. Given a vertex i, let d_i^+ denote the number of successors of i. Note that $\forall i \in [n] \setminus \{1\}$, $d_i^+ = d_i - 1$. Let $(\phi_j)_{j \geq 0}$ denote a sequence of positive integers.

Given a tree $T \in \mathcal{T}$, it is possible to define the weight of a tree, as initially introduced by Meir and Moon [8] as follows:

$$\omega : \mathcal{T} \to \mathbb{N}, \quad \omega(T) = \prod_{i \in T} \phi_{d_i^+} = \prod_{j \geq 0} \phi_j^{N_j(T)},$$

where $N_j(T)$ denotes the number of vertices with exactly j successors in T.

For $n \geq 0$, let $y_n := \sum_{T:|T|=n} \omega(T)$. Note that if every tree T has weight equal to 1, y_n is simply equal to $|\mathcal{T}_n|$.

This allows us to define a probability measure as follows:

$$\forall T \in \mathcal{T}_n, \quad P_n(T) = \frac{\omega(T)}{y_n}.$$

This indeed verifies:

$$\sum_{T \in \mathcal{T}_n} P_n(T) = 1.$$

Remark (URRT model). Define the power series Φ as the exponential generating function of the sequence ϕ_n

$$\Phi(x) := \sum_{n \geq 0} \frac{\phi_n x^n}{n!}$$

and

$$y(z) := \sum_{n \geq 0} \frac{y_n z^n}{n!}.$$

Setting $\phi_j = 1$ for all $j \geq 0$ implies:

- $\forall T$, $\omega(T) = 1$, i.e., every tree has weight equal to one.
- $y_n = |\mathcal{T}_n| = (n-1)!$, as there are n possible ways to attach vertex $n+1$ to a tree of size n.
- Finally, define for any tree $T \in \mathcal{T}_n$

$$P_n(T) := \frac{\omega(T)}{y_n} = \frac{1}{|\mathcal{T}_n|} = \frac{1}{(n-1)!}.$$

P_n is the uniform measure on the set $\mathcal{T}_n$. Attributing a weight of 1 to every tree and picking a tree uniformly at random is equivalent to building the tree recursively, picking each time a vertex $i \in [n]$ uniformly at random to attach vertex $n+1$.

This results in

$$y(z) = \sum_{n\geq 0} \frac{y_n z^n}{n!} = \sum_{n\geq 1} \frac{z^n}{n} = -\ln(1-z),$$

and

$$\Phi(x) = \sum_{n\geq 0} \frac{\phi_n x^n}{n!} = \sum_{n\geq 0} \frac{x^n}{n!} = \exp(x).$$

Remark. Note that in the URRT model, Φ can equivalently be defined as $\Phi(x) = \sum_{n\geq 0} \phi_n x^n$, where $\phi_n = \frac{1}{n!}$ (see Wagner and Durant, Section 2 [11]).

(PA model). Let us define the ordinary generating function of the sequence ϕ_n

$$\Phi(x) := \sum_{n\geq 0} \phi_n x^n$$

in the PA model. Let $\phi_j = 1$ for all $j \geq 0$. This implies $\omega(T) = 1$ for all T, $y_n = |\mathcal{T}'_n| = 1 \cdot 3 \cdot 5 \cdots (2n-3) = (2n-3)!!$, where $\mathcal{T}'_n$ denotes the set of finite rooted trees of size n, where vertices are ordered (i.e., the successors of a vertex are distinguished). Finally, define for any tree $T \in \mathcal{T}'_n$

$$P_n(T) := \frac{\omega(T)}{y_n} = \frac{1}{|\mathcal{T}'_n|} = \frac{1}{(2n-3)!!}.$$

Similarly, P_n is now the uniform measure on the set $\mathcal{T}'_n$: attributing a weight of 1 to every plane-oriented tree (distinguishing the left-to-right successors of a vertex) and picking a tree uniformly at random is equivalent to building the tree iteratively according to the PA model, i.e., attaching a new vertex to existing vertices with a probability proportional to the degree of each vertex. Indeed, distinguishing the successors of a vertex when counting the number of trees in the set $\mathcal{T}'_n$ (which is what is done above, as every tree has weight equal to 1) is equivalent to modelling the preferential attachment rule when iteratively building the tree.

This results in

$$y(z) = \sum_{n\geq 0} \frac{y_n z^n}{n!} = 1 - \sqrt{1-2z}$$

and

$$\Phi(x) = \sum_{n\geq 0} \phi_n x^n = \sum_{n\geq 0} x^n = \frac{1}{1-x}.$$

One can observe from the two examples above that the sequence $(\phi_n)_{n\geq 0}$ encodes the attachment rule, or, said differently, the tree family. These known results, specific to the URRT and PA models, have more general expressions. Given a sequence $(\phi_n)_{n\geq 0}$, let $\Phi(x) = \sum_{n\geq 0} \phi_n x^n$ be its characteristic function. It can be shown (see Wagner and Durant, Section 2 [11]) that the expression of the characteristic function for the general plane-oriented tree (i.e., where successors of a node are distinguished) is

$$\Phi(x) = (1 + c_2 x)^{1+\frac{c_1}{c_2}},$$

with $c_2 < 0$ and $\frac{c_1}{c_2} < -1$. This general expression for the characteristic function is based on a lemma from Panholzer and Prodinger [10], where c_1, c_2 verify $y_{n+1}/y_n = c_1 n + c_2$, with $y_n = \sum_{T:|T|=n} \omega(T)$ being the total weight of trees of size n.

For instance, in the PA model, one can observe that $y_n = (2n-1)!!$ verifies $y_{n+1}/y_n = 2n-1$. And indeed, taking $c_2 = -1$ and $c_1 = 2$ results in

$$\Phi(x) = (1-x)^{-1},$$

which is the result we had previously.

In the affine attachment case, where the weight of vertex i is equal to $d_i + a$, $a \geq 0$, let $\phi_0 = 1$ and $\phi_n = \frac{(a+1)\cdots(a+n)}{n!}$ for all $n \geq 1$. Using the generalized binomial formula, we obtain the expression below for the characteristic function of the affine attachment model:

$$\begin{aligned}\Phi(x) &= \sum_{n\geq 0} \phi_n x^n \\ &= \sum_{n\geq 0} \binom{-1-a}{n} (-1)^n x^n \\ &= (1-x)^{1-a},\end{aligned}$$

which corresponds to taking $c_2 = -1$ and $c_1 = a + 2$ in the general expression. Note this is another way to see that taking $a = 0$ in the affine attachment model coincides with the PA model.

Recall $y_n = \sum_{T:|T|=n} \omega(T)$ and $y(x) = \sum_{n\geq 0} \frac{y_n x^n}{n!}$. Working from the derived power series

$$y'(x) = (1 - (a+2)x)^{-\delta},$$

where $\delta = 1 - \frac{1}{a+2}$, Wagner and Durant [11] show (see Theorems 6 and 9) that the random variable D_n converges in expectation to a limit variable D. They obtain

$$\lim_{n\to\infty} \mathbb{E}[D_n] = \mathbb{E}[D] = \delta.$$

Note $a \geq 0$ implies $\delta \in [1/2, 1)$. This provides us with the useful (asymptotic) upper bound which will serve later on. It holds that $\forall a \geq 0$, for a given n_0 large enough and $\forall n \geq n_0$,

$$\boxed{\mathbb{E}[D_n] \leq \mathbb{E}[D] + 1 \leq \delta + 1 \leq 2.} \tag{1}$$

3 Uniform Attachment Model

3.1 Upper Bound

We consider here the URRT model, and propose an upper bound for $\mathbb{E}\left[|\hat{\sigma}_J(i) - \sigma(i)|\right]$. $\hat{\sigma}_J$ denotes the ordering obtained using the Jordan centrality. Given this estimator is label-invariant, we can assume the labelling of the vertices is their time of arrival (i.e. $\sigma = \mathrm{Id}$). It therefore amounts to upper bounding the quantity $\mathbb{E}\left[|\hat{\sigma}_J(i) - i|\right]$.

Proposition 3.1. *This proposition is a specific case of Proposition* **??***, which provides a result in a more general setting. $D_n = d(1, c)$ is the random variable equal to the distance between the root and the centroid (taking the closest to the root if two centroids exist) in a tree T_n. For any vertex $i \geq 3$, it holds that:*

$$\mathbb{E}[|\hat{\sigma}_J(i) - i|] \leq \mathbb{E}[D_n] + C \cdot i \left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^2\right).$$

Proof. From Lemma 1.1, we obtain that for any vertex $i \in [n]$, $|\hat{\sigma}_J(i) - \hat{\sigma}'(i)| \leq D_n$. Hence,

$$\mathbb{E}[|\hat{\sigma}_J(i) - i|] = \mathbb{E}[|\hat{\sigma}_J(i) - \hat{\sigma}'(i) + \hat{\sigma}'(i) - i|] \leq \mathbb{E}[D_n] + \mathbb{E}[|\hat{\sigma}'(i) - i|]$$

$$\leq \mathbb{E}[D_n] + C \cdot i \left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^2\right).$$

This is a result we prove in a more general setting later in the paper, in Proposition **??**. □

3.2 Lower Bound

Still considering the URRT model, we propose a lower bound for $R_{1,j}(\hat{\sigma}) = \mathbb{E}\left[\frac{|\hat{\sigma}(j) - \sigma(j)|}{\sigma(j)}\right]$, valid for any label-invariant estimator.

Theorem 3.2. *For any label-invariant estimator $\hat{\sigma}$, any permutation σ applied to vertices, noting $\tau = \sigma^{-1}$ the inverse function of σ, and for any $j \in [2, \lfloor \frac{n}{2} \rfloor]$*

$$\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j|\right] + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j|\right] \geq \frac{1}{6} j.$$

Furthermore, a direct consequence is that the maximal risk over all vertices verifies:

$$\max_{i \in [n]} R_{1,i}(\hat{\sigma}) = \max_{i \in [n]} \mathbb{E}\left[\frac{|\hat{\sigma}(i) - \sigma(i)|}{\sigma(i)}\right] \geq \frac{1}{24}\left(1 - \frac{1}{n}\right).$$

Proof. Let us first establish that

$$\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j|\right] + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j|\right] \geq \frac{1}{6} j.$$

Consider a tree T_n with n vertices generated by the URRT model, and an ordering σ assigning the time of arrival to each vertex's label. We note $\tau = \sigma^{-1}$ the inverse function of σ (i.e., τ assigns a label to a time of arrival). For any label-invariant ordering estimator $\hat{\sigma}$, and any vertex $j \in [n]$, we have:

$$\begin{aligned}
R_{1,j}(\hat{\sigma}) &= \mathbb{E}\left[\frac{|\hat{\sigma}(j)-\sigma(j)|}{\sigma(j)}\right] \\
&= \mathbb{E}\left[\frac{|\hat{\sigma}\circ\tau(i)-i|}{i}\right] \quad \text{for a given } i\in[n]\text{, as } \tau \text{ is a bijection} \\
&\geq \frac{1}{n}\mathbb{E}\left[|\hat{\sigma}\circ\tau(i)-i|\right]. \qquad (1)
\end{aligned}$$

$\forall j \geq 2$, let us introduce the event A_j as follows:

$$A_j := \{\text{vertices } j \text{ and } 2j \text{ are leaves, connected to vertices whose arrival time is } \leq j-1\}.$$

Note A_j is the intersection of two independent events, whose probabilities can be directly calculated, namely:

$$A_{j,1} := \{\text{vertices } j+1,\dots,2j-1 \text{ do not connect to vertex } j\},$$

whose probability is

$$\frac{j-1}{j}\times\frac{j}{j+1}\times\cdots\times\frac{2j-3}{2j-2}=\frac{1}{2}, \quad \text{and}$$

$$A_{j,2} := \{\text{vertex } 2j \text{ connects to a vertex whose arrival time is} \leq j-1\},$$

whose probability is

$$\frac{j-1}{2j-1}.$$

By independence, we thus obtain, $\forall j \geq 2, \mathbb{P}\{A_j\} = \frac{j-1}{2(2j-1)} \geq \frac{1}{6}$.

Conditioning on A_j, we can lower bound our initial expression as follows, for any $j \in [2, \lfloor \frac{n}{2} \rfloor]$:

$$\mathbb{E}\left[|\hat{\sigma}\circ\tau(j)-j|+|\hat{\sigma}\circ\tau(2j)-2j|\right] \geq \mathbb{P}\{A_j\}\mathbb{E}\left[|\hat{\sigma}\circ\tau(j)-j|+|\hat{\sigma}\circ\tau(2j)-2j| \mid A_j\right].$$

Let us note δ the permutation swapping vertices j and $2j$ (i.e., δ is the $(j,2j)$ transposition). T^δ denotes the tree T with δ applied to its vertices. Decomposing on the various realizations of a tree T following the URRT model, it holds that:

$$\begin{aligned}
\mathbb{E}\left[|\hat{\sigma}\circ\tau(j)-j|+|\hat{\sigma}\circ\tau(2j)-2j| \mid A_j\right] &= \sum_t \mathbb{E}\left[|\hat{\sigma}\circ\tau(j)-j| \mid A_j, T=t\right]\mathbb{P}\{T=t \mid A_j\} \\
&+ \sum_t \mathbb{E}\left[|\hat{\sigma}\circ\tau(2j)-2j| \mid A_j, T=t^\delta\right]\mathbb{P}\{T=t^\delta \mid A_j\}
\end{aligned}$$

As a consequence of Theorem 4 of Crane and Xu [5], which establishes that two trees T_1 and T_2 generated by the URRT model with the same shape albeit a different labelling have the same probability (the URRT model is said to be shape exchangeable), it follows that $\mathbb{P}\{T = t \mid A_j\} =$

$\mathbb{P}\{T = t^\delta \mid A_j\}$.

The expression above can hence be rewritten as:

$$\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j| + |\hat{\sigma} \circ \tau(2j) - 2j| \mid A_j\right] = \sum_t \mathbb{P}\{T = t \mid A_j\} \left(\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j| \mid A_j, T = t\right] \right.$$
$$\left. + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j| \mid A_j, T = t^\delta\right]\right).$$

Using the triangle inequality and the fact that

$$(\hat{\sigma} \circ \tau(j) \mid A_j, T = t) \stackrel{d}{=} (\hat{\sigma} \circ \tau(2j) \mid A_j, T = t^\delta),$$

we obtain

$$\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j| \mid A_j, T = t\right] + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j| \mid A_j, T = t^\delta\right] \geq j,$$

which, using the conditioning on A_j, results in

$$\boxed{\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j|\right] + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j|\right] \geq \frac{1}{6} j, \quad \text{for any } j \in \left[2, \left\lfloor \frac{n}{2} \right\rfloor\right].}$$

This is the first lower bound we wanted to establish. Finally, evaluating the expression above for $j = \lfloor \frac{n}{2} \rfloor$, it holds that:

$$\mathbb{E}\left[\left|\hat{\sigma} \circ \tau\left(\left\lfloor \frac{n}{2} \right\rfloor\right) - \left\lfloor \frac{n}{2} \right\rfloor\right|\right] + \mathbb{E}\left[\left|\hat{\sigma} \circ \tau\left(2\left\lfloor \frac{n}{2} \right\rfloor\right) - 2\left\lfloor \frac{n}{2} \right\rfloor\right|\right] \geq \frac{1}{6}\left\lfloor \frac{n}{2} \right\rfloor \geq \frac{n-1}{12}.$$

Hence,

$$\mathbb{E}\left[\left|\hat{\sigma} \circ \tau\left(\left\lfloor \frac{n}{2} \right\rfloor\right) - \left\lfloor \frac{n}{2} \right\rfloor\right|\right] \geq \frac{n-1}{24} \quad \text{or} \quad \mathbb{E}\left[\left|\hat{\sigma} \circ \tau\left(2\left\lfloor \frac{n}{2} \right\rfloor\right) - 2\left\lfloor \frac{n}{2} \right\rfloor\right|\right] \geq \frac{n-1}{24}.$$

This implies

$$\max_{j \in [n]} \mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j|\right] \geq \frac{n-1}{24}.$$

which, plugged back into (1), concludes the proof:

$$\boxed{\max_{j \in [n]} R_{1,j}(\hat{\sigma}) \geq \frac{1}{24}\left(1 - \frac{1}{n}\right).}$$

□

4 Affine Attachment Model

4.1 Context

In the affine attachment (AA) model, the attachment of a new node to an existing tree is based on vertex i having a weight of $d_i + a$, where d_i is the degree of vertex i and a is a parameter between 0 and $+\infty$.

Hence, given a tree T_n of size n,

$$\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = \frac{d_i + a}{\sum_i (d_i + a)} = \frac{d_i + a}{2(n-1) + an} = \frac{d_i + a}{(2+a)n - 2}$$

Note that $a = 0$ corresponds to the preferential attachment model, and $a \to \infty$ to the uniform attachment model.

Let T_n be a tree with n vertices. In such a model, for a given tree T_1 with n_1 vertices, we note that:

$$\mathbb{P}(n+1 \text{ connects to } T_1) = \frac{\sum_{i \in T_1} (d_i + a)}{(2+a)n - 2} = \frac{(2+a)n_1 - 1}{(2+a)n - 2}$$

Let $\gamma := \frac{1}{2+a}$. We hence have

$$\mathbb{P}(n+1 \text{ connects to } T_1) = \frac{n_1 - \gamma}{n - 2\gamma}.$$

Thus, $\gamma = \frac{1}{2}$ corresponds to the preferential attachment model, and $\gamma \to 0$ to the uniform attachment model.

For a vertex j, we denote by $\mathrm{de}(j)$ the set of vertices i whereby j belongs to the path between the root and vertex i, excluding j (in other words the strict descendants of j).

We observe that

$$\mathbb{P}(\text{among } m \text{ new connections, } j + i_1, \ldots, j + i_k \in \mathrm{de}(j)) =$$

$$\frac{(j-1-\gamma)}{(j-2\gamma)} \cdot \frac{(j-\gamma)}{(j+1-2\gamma)} \cdot \ldots \cdot \frac{(j+i_1-3-\gamma)}{(j+i_1-2-2\gamma)} \left\{ {}^{|i_1 - 1| \sim T \setminus T_1} \right\}$$

$$\cdot \frac{(1-\gamma)}{(j+i_1-1-2\gamma)} \left\{ {}^{i_1 \sim T_1} \right\}$$

$$\cdot \frac{(j+i_1-2-\gamma)}{(j+i_1-2\gamma)} \left\{ {}^{i_1+1 \sim T \setminus T_1} \right\} \cdot$$

$$\cdot \frac{(2-\gamma)}{(j+i_2-1-2\gamma)} \left\{ {}^{i_2 \sim T_1} \right\} \cdot \ldots \cdot \frac{(k-\gamma)}{(j+i_k-1-2\gamma)} \left\{ {}^{i_k \sim T_1} \right\} \cdot$$

$$\cdot \frac{(j+m-k-2-\gamma)}{(j+m-1-2\gamma)} \left\{ {}^{j+m \sim T \setminus T_1} \right\}$$

$$= \frac{(1-\gamma)(2-\gamma)\ldots(k-\gamma)}{(j-2\gamma)(j+1-2\gamma)\ldots(j+m-1-2\gamma)} \cdot (j-1-\gamma)(j-\gamma)\ldots(j+m-k-2-\gamma)$$

Let us note this probability $p_{j,m,k}$. Indeed, $p_{j,m,k}$ doesn't depend on the values of $i_1, \ldots, i_k$ (i.e., the order of the various connections). With a slight abuse of notation, we identify the set $\mathrm{de}(j)$ with its cardinal, and note $\mathrm{de}_n(j)$ the number of descendants of j at a given time n. We thus obtain that:

$$\begin{aligned}\mathbb{P}(\mathrm{de}_n(j) = k) &= \binom{n-j}{k} \cdot p_{j,n-j,k} \\ &= \binom{n-j}{k} \cdot \frac{(1-\gamma)(2-\gamma)\cdots(k-\gamma)}{(j-2\gamma)(j+1-2\gamma)\cdots(n-1-2\gamma)} \cdot (j-1-\gamma)(j-\gamma)\cdots(n-k-2-\gamma)\end{aligned}$$

4.2 Upper Bound

Given a vertex $i \geq 3$, we now look to upper bound $\mathbb{E}[|\hat{\sigma}_J(i) - i|]$.

We start by upper bounding $\mathbb{E}[\hat{\sigma}'(i)]$ for a given vertex i. Recall $\hat{\sigma}_J$ denotes the estimator ordering vertices using the Jordan centrality, and $\hat{\sigma}'$ denotes the estimator ordering vertices by their number of descendants.

Theorem 4.1. *For any vertex $i \geq 3$, it holds that:*

$$\mathbb{E}[\hat{\sigma}'(i)] \leq i \left[1 + \left(\frac{n}{i}\right)^{\gamma} + \frac{15}{\gamma}\left(\left(\frac{n}{i}\right)^{\gamma} - 1\right)\left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right)\right],$$

which can also be expressed as

$$\mathbb{E}[\hat{\sigma}'(i)] \lesssim n^{\gamma} i^{1-\gamma}\left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right),$$

where the constant on the right hand side also depends on γ. Specifically, in the PA and URRT models, we obtain, respectively:

$$\mathbb{E}[\hat{\sigma}'(i)] \leq C_{1/2} \cdot \sqrt{i \cdot n}\left(1 + \sqrt{\left(\ln\left(\frac{n}{i}\right)\right)}\right), \text{ and}$$

$$\mathbb{E}[\hat{\sigma}'(i)] \leq C_0 \cdot i \left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^2\right).$$

Proof. Fix n, let T_n be a tree of size n, and i a vertex. We have:

$$\hat{\sigma}'(i) \leq n - \#\{j : \mathrm{de}(j) < \mathrm{de}(i)\} = \#\{j : \mathrm{de}(j) \geq \mathrm{de}(i)\}$$

Hence:

$$\begin{aligned}\mathbb{E}[\hat{\sigma}'(i)] &\leq \mathbb{E}\left[\sum_{j \neq i} \mathbb{1}_{\{\mathrm{de}(j) \geq \mathrm{de}(i)\}}\right] + 1 \\ &\leq \sum_{j \neq i} \mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) + 1\end{aligned}$$

We also note that for any $\tau > 0$,

$$\mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) \leq \mathbb{P}\left(\frac{\mathrm{de}(j)}{n} \geq \tau\right) + \mathbb{P}\left(\frac{\mathrm{de}(i)}{n} \leq \tau\right)$$

Upper bounding $\mathbb{P}(\mathrm{de}(j) = k)$ for $j \geq 3$, can be done as follows. We note that:

$$\mathbb{P}(\mathrm{de}(j) = k) = A \cdot B \cdot C$$

where:

$$\begin{aligned} A &= \frac{(1-\gamma)\cdots(k-\gamma)}{k!} \\ B &= \frac{(j-1-\gamma)\cdots(n-2-\gamma)}{(j-2\gamma)\cdots(n-1-2\gamma)} \\ C &= \frac{(n-j)!/(n-j-k)!}{(n-k-1-\gamma)\cdots(n-2-\gamma)} \end{aligned}$$

We have:

$$A = \prod_{l=1}^{k} \frac{l-\gamma}{l}$$

We can rewrite A as:

$$A = \exp\left(\sum_{l=1}^{k} \ln\left(\frac{l-\gamma}{l}\right)\right) \leq \exp\left(-\gamma \sum_{l=1}^{k} \frac{1}{l}\right)$$

Which results in:

$$\boxed{A \leq \frac{1}{k^\gamma}}$$

Next, for $j \geq 2$,

$$\begin{aligned} B &= \prod_{l=j}^{n-1} \frac{l-1-\gamma}{l-2\gamma} \\ &= \prod_{l=j}^{n-1} \left(1 + \frac{\gamma-1}{l-2\gamma}\right) \\ &\leq \exp\left((\gamma-1)\sum_{l=j}^{n-1} \frac{1}{l-2\gamma}\right) \\ &\leq \exp\left((\gamma-1)\sum_{l=j}^{n-1} \frac{1}{l}\right) \\ &\leq \exp\left((\gamma-1)\int_j^n \frac{dt}{t}\right) \\ &= \left[\exp\left(\ln\left(\frac{n}{j}\right)\right)\right]^{\gamma-1} \\ &= \left(\frac{n}{j}\right)^{\gamma-1}. \end{aligned}$$

And hence:

$$\boxed{B \leq \left(\frac{n}{j}\right)^{\gamma-1}}$$

Finally, for $j \geq 3$, which implies $j - 2 - \gamma \geq 0$,

$$\begin{aligned}
C &= \frac{(n-j-k+1)\cdots(n-j)}{(n-k-1-\gamma)\cdots(n-2-\gamma)} \\
&= \prod_{l=n-k+1}^{n} \frac{l-j}{l-2-\gamma} \\
&= \prod_{l=n-k+1}^{n} \left(1 + \frac{2+\gamma-j}{l-2-\gamma}\right) \\
&= \exp\left(\sum_{l=n-k+1}^{n} \ln\left(1 + \frac{2+\gamma-j}{l-2-\gamma}\right)\right) \\
&\leq \exp\left(\sum_{l=n-k+1}^{n} \frac{2+\gamma-j}{l-2-\gamma}\right) \\
&= \exp\left(\sum_{l=n-k-1}^{n-2} \frac{2+\gamma-j}{l-\gamma}\right) \\
&\leq \exp\left((2+\gamma-j) \sum_{l=n-k-1}^{n-2} \frac{1}{l}\right) \\
&\leq \exp\left((2+\gamma-j) \int_{n-k-1}^{n-1} \frac{dt}{t}\right) \\
&= \left(\frac{n-1}{n-k-1}\right)^{2+\gamma-j} \\
&= \left(\frac{n-k-1}{n-1}\right)^{j-2-\gamma} \\
&\leq \left(\frac{n-k}{n}\right)^{j-2-\gamma} \\
&= \left(1 - \frac{k}{n}\right)^{j-2-\gamma} \\
&\leq \left(1 - \frac{k}{n}\right)^{j-3}.
\end{aligned}$$

Hence,

$$\boxed{C \leq \left(1 - \frac{k}{n}\right)^{j-3}}$$

Thus, for $k \geq 1$ and $j \geq 3$,

$$\mathbb{P}(\mathrm{de}_n(j) = k) \leq \frac{1}{k^\gamma} \cdot \left(\frac{n}{j}\right)^{\gamma-1} \cdot \left(1 - \frac{k}{n}\right)^{j-3}$$

For $k = 0$ and $j \geq 3$,

$$\mathbb{P}(\text{de}(j) = 0) = \mathbb{P}(j \text{ has no descendants}) = \frac{(j-1-\gamma)(j-\gamma)\cdots(n-2-\gamma)}{(j-2\gamma)(j+1-2\gamma)\cdots(n-1-2\gamma)} = B \leq \left(\frac{n}{j}\right)^{\gamma-1}.$$

Hence, for any $\tau > 0$ and for any vertex $i \geq 3$,

$$\begin{aligned}
\mathbb{P}\left(\frac{\text{de}(i)}{n} \leq \tau\right) &= \mathbb{P}(\text{de}(i) = 0) + \sum_{k=1}^{\lfloor \tau n \rfloor} \mathbb{P}(\text{de}(i) = k) \\
&\leq \left(\frac{n}{i}\right)^{\gamma-1} + \sum_{k=1}^{\lfloor \tau n \rfloor} \frac{1}{k^\gamma} \cdot \left(\frac{n}{i}\right)^{\gamma-1} \cdot \left(1 - \frac{k}{n}\right)^{i-3} \\
&\leq \left(\frac{n}{i}\right)^{\gamma-1} \left[1 + \sum_{k=1}^{\lfloor \tau n \rfloor} \frac{1}{k^\gamma}\right].
\end{aligned}$$

Given

$$\sum_{k=1}^{\lfloor \tau n \rfloor} \frac{1}{k^\gamma} \leq \frac{(\tau n)^{1-\gamma}}{1-\gamma},$$

it holds that

$$\begin{aligned}
\mathbb{P}\left(\frac{\text{de}(i)}{n} \leq \tau\right) &\leq \left(\frac{n}{i}\right)^{\gamma-1} \left[1 + \frac{(\tau n)^{1-\gamma}}{1-\gamma}\right] \\
&\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{(\tau i)^{1-\gamma}}{(1-\gamma)}.
\end{aligned}$$

Similarly, $\forall j \geq 3$,

$$\begin{aligned}
\mathbb{P}\left(\frac{\text{de}(j)}{n} \geq \tau\right) &\leq \sum_{k=\lfloor \tau n \rfloor}^{n} \mathbb{P}(\text{de}(j) = k) \\
&\leq \sum_{k=\lfloor \tau n \rfloor}^{n} \frac{1}{k^\gamma} \cdot \left(\frac{n}{j}\right)^{\gamma-1} \cdot \left(1 - \frac{k}{n}\right)^{j-3} \\
&\leq \left(\frac{n}{j}\right)^{\gamma-1} \cdot \frac{1}{\lfloor \tau n \rfloor^\gamma} \cdot \sum_{k=\lfloor \tau n \rfloor}^{n} \left(1 - \frac{k}{n}\right)^{j-3}.
\end{aligned}$$

Using a sum/integral comparison,

$$\begin{aligned}
\frac{1}{n} \sum_{k=\lfloor \tau n \rfloor}^{n} \left(1 - \frac{k}{n}\right)^{j-3} &\leq \sum_{k=\lfloor \tau n \rfloor}^{n} \int_{\frac{k-1}{n}}^{\frac{k}{n}} (1-t)^{j-3}\, dt \\
&= \int_{\frac{\lfloor \tau n \rfloor - 1}{n}}^{1} (1-t)^{j-3}\, dt \\
&= \frac{\left(1 - \frac{\lfloor \tau n \rfloor}{n} + \frac{1}{n}\right)^{j-2}}{j-2} \\
&\leq \frac{\left(1 - \tau + \frac{2}{n}\right)^{j-2}}{j-2}.
\end{aligned}$$

Furthermore, upper bounding $\frac{n}{\lfloor \tau n \rfloor}$ can be done as follows:

$$\frac{n}{\lfloor \tau n \rfloor} \leq \frac{n}{\tau n - 1} = \frac{n}{\tau(n - \frac{1}{\tau})}.$$

Note that $\frac{n}{n-\frac{1}{\tau}} \leq C$ if and only if:

$$\begin{aligned} n &\leq Cn - \frac{C}{\tau} \\ \iff \frac{C}{\tau} &\leq n(C-1) \\ \iff \frac{\tau}{C} &\geq \frac{1}{n(C-1)} \\ \iff \tau &\geq \frac{C}{n(C-1)}. \end{aligned}$$

Hence, provided $\tau \geq \frac{2}{n}$, it holds that $\frac{n}{n-\frac{1}{\tau}} \leq 2$, and thus:

$$\frac{n}{\lfloor \tau n \rfloor} \leq \frac{2}{\tau}.$$

We finally obtain:

$$\mathbb{P}\left(\frac{\mathrm{de}(j)}{n} \geq \tau\right) \leq 2^{\gamma} \cdot \frac{1}{j^{\gamma-1}} \cdot \frac{1}{\tau^{\gamma}} \cdot \frac{\left(1-\tau+\frac{2}{n}\right)^{j-2}}{j-2}.$$

This results in the following upper bound: $\forall \tau \geq \frac{2}{n}, \forall i, j \geq 3$,

$$\mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) \leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{(\tau i)^{1-\gamma}}{(1-\gamma)} + 2^{\gamma} \cdot \frac{1}{j^{\gamma-1}} \cdot \frac{1}{\tau^{\gamma}} \cdot \frac{\left(1-\tau+\frac{2}{n}\right)^{j-2}}{j-2}.$$

Substituting τ for $\frac{l}{j-2}$, where $l := 2 + (1-\gamma)\ln\left(\frac{j}{i}\right)$ and $j > i$, we obtain the following set of inequalities:

$$\begin{aligned} \mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{il}{(j-2)}\right)^{1-\gamma} + 2^{\gamma}\left(\frac{jl}{j-2}\right)^{1-\gamma}\left(\frac{j-2}{l}\right)\frac{\left(1-\tau+\frac{2}{n}\right)^{j-2}}{j-2} \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{i}{j-2}\right)^{1-\gamma} l^{1-\gamma} + \frac{2^{\gamma}}{l} \cdot 3^{1-\gamma} \cdot l^{1-\gamma}\left(1-\tau+\frac{2}{n}\right)^{j-2} \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{i}{j-2}\right)^{1-\gamma} l^{1-\gamma} + 3 \cdot l^{-\gamma}\exp\left((j-2)\ln\left(1-\tau+\frac{2}{n}\right)\right) \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{i}{j-2}\right)^{1-\gamma} l^{1-\gamma} + 3 \cdot l^{-\gamma}\exp\left((j-2)\left(\frac{2}{n}-\tau\right)\right) \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{i}{j-2}\right)^{1-\gamma} l^{1-\gamma} + 3 \cdot l^{-\gamma} \cdot e^{2} \cdot e^{-l} \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + \frac{1}{1-\gamma}\left(\frac{i}{j}\right)^{1-\gamma}\left(\frac{j}{j-2}\right)^{1-\gamma} l^{1-\gamma} + 3 \cdot l^{-\gamma} \cdot \left(\frac{i}{j}\right)^{1-\gamma} \quad \text{as } e^{-l} = e^{-2}\left(\frac{i}{j}\right)^{1-\gamma} \\ &\leq \left(\frac{n}{i}\right)^{\gamma-1} + 6 \cdot \left(\frac{i}{j}\right)^{1-\gamma}\left(l^{1-\gamma} + \frac{1}{2}l^{-\gamma}\right) \quad \text{as } \frac{1}{1-\gamma}\left(\frac{j}{j-2}\right)^{1-\gamma} \leq 6. \end{aligned}$$

Let us now upper bound $l^{1-\gamma} + \frac{1}{2}l^{-\gamma}$.

$$\text{As } \frac{l}{2} \geq 1, \text{ we obtain } l^{1-\gamma} + \frac{1}{2}l^{-\gamma} \leq \frac{5}{4}l^{1-\gamma}$$
$$\leq \frac{5}{4}\left(2 + (1-\gamma)\ln\left(\frac{j}{i}\right)\right)^{1-\gamma}$$
$$\leq \frac{5}{2}\left(1 + \ln\left(\frac{j}{i}\right)\right)^{1-\gamma}$$
$$\leq \frac{5}{2}\left(1 + \left(\ln\left(\frac{j}{i}\right)\right)^{1-\gamma}\right) \quad \text{as } 0 < 1-\gamma < 1.$$

Therefore, for $i, j \geq 3$,

$$\boxed{\mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) \leq \left(\frac{n}{i}\right)^{\gamma-1} + 15 \cdot \left(\frac{i}{j}\right)^{1-\gamma}\left(1 + \left(\ln\left(\frac{j}{i}\right)\right)^{1-\gamma}\right)}$$

Remark (limit cases). Note that $\gamma \to 0$ (i.e., the URRT model) results in $\tau = \frac{2+\ln(j/i)}{j-2}$, and

$$\mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) \leq \frac{i}{n} + 15\frac{i}{j}\left[1 + \ln\left(\frac{j}{i}\right)\right],$$

and $\gamma = \frac{1}{2}$ (i.e., the PA model), results in $\tau = \frac{4+\ln(j/i)}{2(j-2)}$ and

$$\mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) \leq \sqrt{\frac{i}{n}} + 15\sqrt{\frac{i}{j}}\left[1 + \sqrt{\ln\left(\frac{j}{i}\right)}\right].$$

We can now conclude. For any vertex $i \geq 3$,

$$\mathbb{E}[\hat{\sigma}'(i)] \leq \sum_{\substack{j=1\\ j\neq i}}^{n} \mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) + 1$$
$$\leq \sum_{j=1}^{i-1} \mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) + \sum_{j=i+1}^{n} \mathbb{P}(\mathrm{de}(j) \geq \mathrm{de}(i)) + 1$$
$$\leq i + \sum_{j=i+1}^{n} \left(\frac{n}{i}\right)^{\gamma-1} + 15\left(\frac{i}{j}\right)^{1-\gamma}\left[1 + \left(\ln\left(\frac{j}{i}\right)\right)^{1-\gamma}\right]$$

Using a sum/integral comparison, we observe

$$\sum_{j=i+1}^{n}\left(\frac{1}{j}\right)^{1-\gamma} \leq \int_i^n \frac{dt}{t^{1-\gamma}} = \frac{n^\gamma - i^\gamma}{\gamma}.$$

Furthermore,

$$\begin{aligned}\sum_{j=i+1}^{n}\left(\frac{1}{j}\right)^{1-\gamma}\left(\ln\left(\frac{j}{i}\right)\right)^{1-\gamma} &\leq \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\sum_{j=i+1}^{n}\left(\frac{1}{j}\right)^{1-\gamma} \\ &\leq \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\frac{n^\gamma - i^\gamma}{\gamma}.\end{aligned}$$

Hence, noting that $(n-i)\left(\frac{n}{i}\right)^{\gamma-1} \leq i\left(\frac{n}{i}\right)^{\gamma}$, we obtain

$$\begin{aligned}\mathbb{E}[\hat{\sigma}'(i)] &\leq i + (n-i)\left(\frac{n}{i}\right)^{\gamma-1} + 15\left[i^{1-\gamma}\cdot\frac{1}{\gamma}\left[n^\gamma - i^\gamma\right] + i^{1-\gamma}\cdot\frac{1}{\gamma}\left[n^\gamma - i^\gamma\right]\left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right] \\ &\leq i\left[1 + \left(\frac{n}{i}\right)^{\gamma} + \frac{15}{\gamma}\left(\left(\frac{n}{i}\right)^{\gamma} - 1\right) + \frac{15}{\gamma}\left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\left[\left(\frac{n}{i}\right)^{\gamma} - 1\right]\right] \\ &\leq \boxed{i\left[1 + \left(\frac{n}{i}\right)^{\gamma} + \frac{15}{\gamma}\left(\left(\frac{n}{i}\right)^{\gamma} - 1\right)\left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right)\right]}\end{aligned}$$

which concludes the proof.

□

A direct consequence of the proposition above is the following upper bound:

$$\boxed{\mathbb{E}[|\hat{\sigma}'(i) - i|] \leq \mathbb{E}[\hat{\sigma}'(i)] + i \lesssim n^\gamma i^{1-\gamma}\left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right)}$$

Remark (affine parameter a). In the affine attachment model, vertex i has a weight equal to $d_i + a$, where d_i is the degree of i, with $a \geq 0$. One may legitimately wonder if this model is still well-defined, and whether results on $\mathbb{E}[\hat{\sigma}'(i)]$ still hold if $a < 0$.

Given a tree T_n of size n, recall vertex $n+1$ connects to the tree according to the following distribution: $\forall i \in \{1, \ldots, n\}$,

$$\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = \frac{d_i + a}{(2+a)n - 2}.$$

Hence, for the model to be well-defined, it must at least verify the conditions below:

- $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) \geq 0$,
- $\sum_{i=1}^{n}\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = 1$, which in turn implies $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) \leq 1$.

If $-1 < a < 0$, these conditions are satisfied, the model is well-defined, and we have $\gamma = \frac{1}{2+a} \in (0,1)$. Notably, the upper bound for $\mathbb{E}[\hat{\sigma}'(i)]$ is still valid, as it is currently written, for $\gamma \in (0,1)$.

If $a = -1$, note a tree of size 2 (consisting of 2 vertices and an edge) cannot grow further using the distribution above. Indeed, vertex 3 connects to both vertex 1 and vertex 2 with probability equal to 0, which doesn't satisfy the condition $\sum_{i=1}^{n}\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = 1$.

However, if the starting point is a larger tree where the set $\{i : d_i > 1\}$ is non-empty, the recursive process is well defined for $a = -1$. This would imply the affine parameter a changes in time, i.e., is a function of n, and is a case we do not explore here.

If $-n < a < -1$, the model cannot be defined the way it currently is, as $\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i)$ would potentially take negative values. Indeed,

$$\begin{aligned}\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) &= \frac{d_i + a}{(2+a)n - 2} < 0 \\ &\iff \left(a < -d_i \text{ and } a > \frac{2}{n} - 2\right) \text{ or } \left(a > -d_i \text{ and } a < \frac{2}{n} - 2\right).\end{aligned}$$

For instance, for a large n and taking $-2 < a < -1$, the probability that a new vertex connects to a leaf, whose degree is equal to one, cannot be defined as it would take a negative value. Simply changing the distribution to

$$\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = \frac{\max(0, d_i + a)}{(2+a)n - 2}$$

doesn't solve the issue, as in some pathological cases this probability can be greater than 1: consider for instance a tree with 6 vertices, where vertex 1 is connected to all vertices 2 to 6, and parameter $a = -\frac{3}{2}$, one would obtain

$$\mathbb{P}(\text{vertex } 7 \text{ connects to vertex } 1) = \frac{5 - \frac{3}{2}}{(2 - \frac{3}{2}) \times 6 - 2} = \frac{7}{2} > 1.$$

More generally, for a given n and $a \in (-n, -1)$, if vertex i is such that $d_i > (n-1)(a+2) \iff d_i + a > (2+a)n - 2 \iff \frac{d_i + a}{(2+a)n-2} > 1$, the model is undefined.

Finally, if $a \leq -n$, the recursive process can be defined in the sense that the conditions

- $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) \geq 0$,
- $\forall i \in \{1, \ldots, n\}$, $\mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) \leq 1$,
- $\sum_{i=1}^{n} \mathbb{P}(\text{vertex } n+1 \text{ connects to vertex } i) = 1$

are met.

However, the recursive process would only be licit for $n \leq -a$. Intuituvely, this corresponds to a process limited in time. For a large n, this would correspond to $\gamma < 0$ albeit close to 0. The upper bound for $\mathbb{E}[\hat{\sigma}'(i)]$ would need to be adapted as the current proof does not hold for $\gamma < 0$.

Using the previous proposition, we are able to derive an upper bound for $\mathbb{E}[|\hat{\sigma}_J(i) - i|]$.

Theorem 4.2. *Recall $D_n = d(1, c)$ is the random variable equal to the distance between the root and the centroid (taking the closest to the root if two centroids exist) in a tree T_n. For any vertex $i \geq 3$, it holds that:*

$$\mathbb{E}[|\hat{\sigma}_J(i) - i|] \leq \mathbb{E}[D_n] + C_\gamma \cdot n^\gamma i^{1-\gamma} \left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right).$$

Proof. Using Lemma 1.1, as ψ_T and ψ'_T coincide outside of $\{1 \to c\}$, it holds that for any vertex $i \in [n]$, $|\hat{\sigma}_J(i) - \hat{\sigma}'(i)| \leq D_n$. Hence,

$$\begin{aligned}\mathbb{E}[|\hat{\sigma}_J(i) - i|] &= \mathbb{E}[|\hat{\sigma}_J(i) - \hat{\sigma}'(i) + \hat{\sigma}'(i) - i|] \leq \mathbb{E}[D_n] + \mathbb{E}[|\hat{\sigma}'(i) - i|] \\ &\leq \mathbb{E}[D_n] + C_\gamma \cdot n^\gamma i^{1-\gamma} \left(1 + \left(\ln\left(\frac{n}{i}\right)\right)^{1-\gamma}\right)\end{aligned}$$

□

4.3 Lower Bound

Still considering the affine attachment model, whereby vertex i has a weight of $d_i + a$, $a \in [0, \infty)$, and for any label-invariant estimator $\hat{\sigma}$, we derive a lower bound for $\mathbb{E}[|\hat{\sigma} \circ \tau(j) - j| + |\hat{\sigma} \circ \tau(2j) - 2j|]$.

Theorem 4.3. *In the affine attachment model, for any label-invariant estimator $\hat{\sigma}$ and any vertex $j \in [2, \lfloor n/2 \rfloor]$, it holds that:*

$$\mathbb{E}[|\hat{\sigma} \circ \tau(j) - j| + |\hat{\sigma} \circ \tau(2j) - 2j|] \geq \frac{1}{6}j.$$

Proof. The proof is directly adapted from Proposition **??** where the only meaningful change is related to the attachment process, hence related to the probability of the event A_j used in the conditioning.

Let $A_{j,1}$ denote the event {vertices $j+1, \ldots, 2j-1$ do not connect to vertex j, i.e., j remains a leaf}.

Its probability can be directly computed and is equal to:

$$\frac{(j-1-\gamma)}{(j-2\gamma)} \cdot \frac{(j-\gamma)}{(j+1-2\gamma)} \cdot \ldots \cdot \frac{(2j-3-\gamma)}{(2j-2-2\gamma)}$$

Let $A_{j,2}$ denote the event {vertex $2j$ connects to a vertex of rank $\leq j-1$}. Let T_1 denote the subtree consisting of vertices $\{1, \ldots, j-1\}$. Given that at least vertices j and $j+1$ are connected to T_1, it holds that $\sum_{i \in T_1}(d_i + a) = \sum_{i \in T_1} d_i + a(j-1) \geq 2(j-2) + 1 + 1 + a(j-1) = (j-1)(2+a)$.

As $\gamma = \frac{1}{(2+a)}$, we thus obtain

$$\mathbb{P}(2j \sim T_1) \geq \frac{(j-1)(2+a)}{(2j-1)(2+a)-2} = \frac{(j-1)}{(2j-1-2\gamma)}$$

Finally, noting $A_j = A_{j,1} \cap A_{j,2}$, we have:

$$\begin{aligned}\mathbb{P}(A_j) &= \mathbb{P}(A_{j,1}) \cdot \mathbb{P}(A_{j,2}) \text{ by independence} \\ &\geq \frac{(j-1-\gamma)}{(j-2\gamma)} \cdot \frac{(j-\gamma)}{(j+1-2\gamma)} \cdot \ldots \cdot \frac{(2j-3-\gamma)}{(2j-2-2\gamma)} \cdot \frac{(j-1)}{(2j-1-2\gamma)} \\ &\geq \frac{(j-1-\gamma)}{(j-2\gamma)} \cdot \frac{(j-2\gamma)}{(j+1-2\gamma)} \cdot \ldots \cdot \frac{(2j-3-2\gamma)}{(2j-2-2\gamma)} \cdot \frac{(j-1)}{(2j-1-2\gamma)} \\ &\geq \frac{1}{2} \cdot \frac{(j-1)}{(2j-1-2\gamma)} \\ &\geq \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}\end{aligned}$$

The remaining steps of the proof, i.e., conditioning on A_j, introducing the transposition $\delta = (j, 2j)$, using Theorem 4 of Crane and Xu [5], which establishes that in the URRT, PA or AA model, two trees with the same shape but different labelling have the same probability (these models are said to be shape exchangeable), and using the distribution equality $(\hat{\sigma} \circ \tau(j) \mid A_j, T = t) \stackrel{d}{=} (\hat{\sigma} \circ \tau(2j) \mid A_j, T = t^\delta)$ are exactly the same and still hold in the case of the affine attachment model.

Hence, the initial result from the URRT model (where $\gamma \to 0$) can be extended here for any $\gamma \in (0, 1/2]$ and we obtain:

$$\boxed{\mathbb{E}\left[|\hat{\sigma} \circ \tau(j) - j|\right] + \mathbb{E}\left[|\hat{\sigma} \circ \tau(2j) - 2j|\right] \geq \frac{1}{6} j, \quad \text{for any } j \in \left[2, \left\lfloor \frac{n}{2} \right\rfloor\right].}$$

□

References

[1] R. Abraham and J.-F. Delmas. An introduction to galton-watson trees and their local limits. *arXiv preprint arXiv:1506.05571*, 2015. Available at `https://arxiv.org/abs/1506.05571`.

[2] David Aldous. The continuum random tree. i. *Annals of Probability*, 19(1), 1991.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286, 1999.

[4] Simon Briend, Christophe Giraud, Gàbor Lugosi, and Deborah Sulem. Estimating the history of a random recursive tree. *arXiv preprint arXiv:2403.09755*, 2024. Available at `https://arxiv.org/pdf/2403.09755`.

[5] Harry Crane and Min Xu. Inference on the history of a randomly growing tree. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(4):639–668, 2021.

[6] Michael Drmota. *Random Trees: An Interplay between Combinatorics and Probability*. Springer Vienna, 2009.

[7] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.

[8] A. Meir and J. W. Moon. On the altitude of nodes in random trees. *Canad. J. Math.*, 30(5):997–1015, 1978.

[9] J. Neveu. Arbres et processus de galton-watson. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 22, 1986.

[10] Panholzer and Prodinger. Level of nodes in increasing trees revisited. *Random Structures & Algorithms*, 31(2):203–226, 2007.

[11] Stephan Wagner and Kevin Durant. On the centroid of increasing trees. *Discrete Mathematics & Theoretical Computer Science*, 21, 2019.