Probability Bounds on Persistent Centroid Appearance in Random Growing Trees

Ryan Hu

St. John's College

October 14, 2024

Random Growing Trees

- Network archaeology (Haigh, 1970; Shah & Zaman, 2011; Bubeck, Devroye & Lugosi, 2017)

Definition (Uniform Attachment Model)

A family of random trees (i.e. family of rv's whose values are trees) $\{T_n\}_{n\geq 1}$ is said to follow a uniform attachment model if $|T_n| = n$, $V(T_n) \subseteq V(T_{n+1})$ for all $n \in \mathbb{N}$, and, if we write $V(T_n) = \{v_1, \ldots, v_n\}$, then $\mathbb{P}((v_i, v_{n+1}) \in E(T_{n+1})) = 1/n$ for each $i \leq n$. Also, each time the attachment choices should be independent from the previous ones.

Centrality of Vertices in a Tree

Definition (Centrality)

Given a finite tree T and a vertex $v \in V(T)$. Consider the forest formed by deleting this vertex. The centrality of this vertex is then

$$\psi_T(v) := \text{The Size of the Largest Tree in the Forest Formed.}$$

Definition (Centroid)

Given a finite tree T and a $v \in V(T)$. v is called a centroid of T if it minimises $\psi_T(v)$.

Persistence of Centroid in Uniform Attachment

Theorem (Loh, Jog, 2016)

Let (T_n) be a family of random growing trees given by the uniform attachment. Then, with probability 1, there exists a $N \in \mathbb{N}$ and a $v \in \bigcup_n V(T_n)$ such that for all $n \geq N$, v is the centroid of T_n.

Probability Bounds on Persistent Centroid Appearance

Definition

Let (T_n) be a family of random growing trees given by uniform attachment. We define the random variable

$$T := \min\{N \in \mathbb{N} : \text{there exists } k \in \mathbb{N} \text{ s.t. } \mathcal{C}(T_n) = \{v_k\} \text{ for all } n \geq N\}.$$

- Note that by the above theorem, this random variable is almost surely well-defined.

Probability Bounds on Persistent Centroid Appearance

Goal

Our goal is to find an explicit expression $N(\varepsilon)$ such that

$$\mathbb{P}(T > N(\varepsilon)) \leq \varepsilon.$$

Understanding Centrality Function

Lemma

Let T be a finite tree. Then T has at most 2 centroids. And, if T has 2 centroids, then these 2 centroids are connected by an edge.

Analysis of the Event $\{T > N\}$

- Start by noting that, whenever there is a change in centroid, there must be a stage in between at which there's two centroids. Hence,

$$\{T > N\} = \bigcup_{k=N}^{\infty} \{|\mathcal{C}(T_k)| = 2\}.$$

- The union bound of the above is not enough, instead, consider the union bound given by

$$\{T > N\} = \bigcup_{j=2}^{\infty} \{\text{there exists } k \geq N, i < j \text{ s.t. } \mathcal{C}(T_k) = \{v_i, v_j\}\}.$$

- It's easy to track $(\psi_{T_k}(v_i), \psi_{T_k}(v_j))$ as k evolves, it's just a Polya urn!

de Finetti's Theorem

Definition

A sequence of random variables (X_i) is said to be exchangeable if $(X_1, \ldots, X_n) \overset{d}{=} (X_{\pi(1)} \ldots, x_{\pi(n)})$ for all $n \in \mathbb{N}$.

Theorem (de Finetti, 1931)

A sequence of Bernoulli random variables $(X_i)_{i \geq 1}$ is exchangeable if and only if there exists a distribution function F on $[0, 1]$ such that for all $n \in \mathbb{N}$,

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \int_0^1 \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \, dF(\theta).$$

de Finetti's Theorem

Definition (Polya Urn)

A Polya urn puts x white and y black balls into an urn. At each step, one ball is drawn uniformly at random from the urn, and its color observed; it is then returned in the urn, and an additional ball of the same color is added to the urn.

Theorem (Janson, 2019)

Let (X_n, Y_n) be the Polya urn model with $X_0 = a$, $Y_0 = b$. Then, there exists a rv $W \sim \text{beta}(a, b)$ with $\frac{X_n}{X_n+Y_n} \to W$ almost surely. And, for each $w \in [0,1]$, conditioning on $W = w$, $X_n = X_0 + \sum_{k=1}^{n} Q_k$ with $Q_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(w)$.

Law of Iterated Logarithm Bounds

- Law of Iterated Logarithm is a classical result in probability theory. It says

$$\limsup_{n\to\infty} \frac{|\sum_{k=1}^n Y_k|}{\sqrt{2n \log\log n}} = 1 \text{ a.s.},$$

  where (Y_k) are i.i.d. rvs with zero mean and unit variances.
- Recent results in sequential hypothesis testing use a non-asymptotic result (Balasubramani & Ramdas, 2016):

Theorem

Let $Q_k \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(w)$, $\eta_t = \eta_0 + \sum_{k=1}^t (-1)^{Q_k}$. Then, with probability $\geq 1-\delta$, for all $t \geq \frac{C_1}{w(1-w)} \log(4/\delta)$ uniformly, we have

$$|\eta_t - (2w-1)t| \leq \sqrt{C_2 w(1-w)t(2\log\log(C_3 w(1-w)t/|\eta_t|) + \log(2/\delta))},$$

where C_1, C_2, C_3 are some explicit constants.

Conclusion

- Analyzed time to persistence for centroid in uniform attachment trees.
- Results rely on convergence in Polya urns (de Finetti's Theorem), finite-time Law of the Iterated Logarithm.
- Some details still to be worked out.
- Future work: Preferential attachment trees, diffusion in d-regular trees.