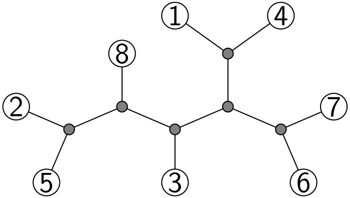
On the Maximum Agreement Subtree of random trees

Thomas Budzinski (CNRS and ENS de Lyon)

2023, March 13th
Journées ALEA 2023
joint work with Delphin Sénizergues (Nanterre)

Labelled binary trees

- A *binary tree* is a finite tree where all vertices have degree either 1 (leaves) or 3 (nodes).
- We consider *labelled binary trees*, i.e. binary trees with n leaves labelled from 1 to n.

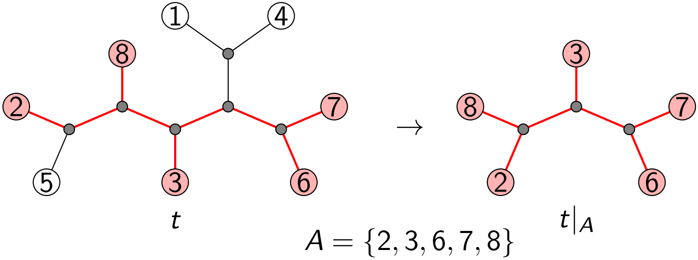


- Simple combinatorial structure: we pass from $n-1$ to n by grafting the leaf n on one of the $2n-5$ edges, so

$$\#\mathcal{T}_n = (2n-5)!! = 1 \times 3 \times 5 \times \cdots \times (2n-5).$$

Subtrees

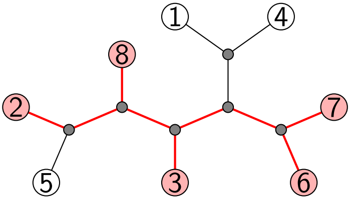
- Let t be a labelled binary tree with n leaves, and A a subset of $\{1, \dots, n\}$. The *subtree of t induced by A* is the labelled binary tree formed by the leaves of t whose label belong to A, and the branches between them.



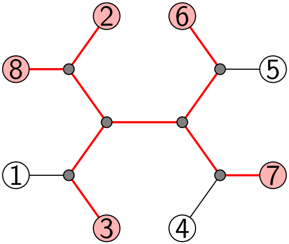
- Maximum Agreement Subtree: if t, t' are labelled binary trees of size n, we write

$$\mathrm{MAST}(t, t') = \max\{|A| \text{ such that } t|_A = t'|_A\}.$$

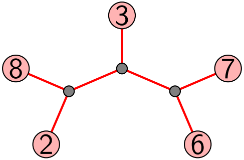
Maximum Agreement Subtree: an example



t



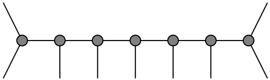
t'



$$\mathrm{MAST}(t, t') = 5$$

Maximum Agreement Subtree

- Motivations:
  - When two different phylogeny methods give different results, measure by how much they disagree and how much information can be saved.
  - Generalization of the longest monotone subsequence of a permutation, when both trees are caterpillars:



- First results:
  - Computation: simple quadratic algorithm **[Steel–Warnow 93]**, improved to $O(n \log n)$ **[Cole–Farach–Hariharan–Przytycka–Thorup 00]**.
  - Worst case **[Markin 18, Kubicka–Kubicki–Morris 92]**:

$$c \log n \leq \min_{|t|=|t'|=n} \mathrm{MAST}(t, t') \leq C \log n.$$

MAST of random trees

- Let T_n, T'_n be two independent labelled binary trees of size n, picked uniformly at random. Order of magnitude of $\mathrm{MAST}(T_n, T'_n)$?
- Motivation: it should not be the case on "real" data, but gives a benchmark.
- First moment upper bound [Bryant–McKenzie–Steel 03]:

$$\mathbb{P}\left(\mathrm{MAST}(T_n, T'_n) \geq k\right) \leq \sum_{\substack{A\subset\{1,\dots,n\},\,|A|=k \\ t \text{ labelled by } A}} \mathbb{P}\left(T_n|_A = T'_n|_A = t\right)$$

$$= \binom{n}{k} \times (2k-5)!! \times \frac{1}{(2k-5)!!^2},$$

since the restriction of T_n to any subset A is uniform. By Stirling, we find $\mathrm{MAST}(T_n, T'_n) = O(\sqrt{n})$ with high probability.

MAST of random trees: lower bounds and particular cases

- Polynomial lower bound: $\mathrm{MAST}(T_n, T'_n) \geq n^{1/8}$ by finding a common caterpillar **[Bernstein–Ho–Long–Steel–St. John–Sullivant 15]**.
- Lower bound increased to $n^{\frac{\sqrt{3}-1}{2}} \approx n^{0,366}$ **[Aldous 20]** and then to $n^{0,4464}$ **[Khezeli 22]**.
- If both trees T_n and T'_n are caterpillars, $\mathrm{MAST}(T_n, T'_n)$ is the length of the longest monotone subsequence of a uniform permutation, so it is $\approx \sqrt{n}$.
- If T_n and T'_n are conditionned to have the same shape (i.e. independent labellings of the same tree t), then $\mathrm{MAST}(T_n, T'_n) \approx \sqrt{n}$ **[Misra–Sullivant 19]**:
    - Divide t into $\sqrt{n}$ regions (R_i) of size $\approx \sqrt{n}$, and take one well chosen label for each region.

MAST of random trees: improved upper bound

Theorem (B.–Sénizergues 23+)

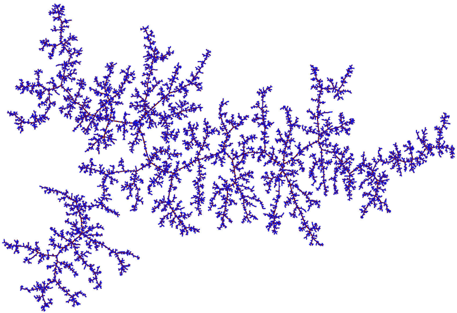
There is $\varepsilon > 0$ such that, with probability $1 - o(1)$, we have

$$\mathrm{MAST}(T_n, T_n') \leq n^{1/2-\varepsilon}.$$

- Explicit ε: very bad ($\varepsilon = 10^{-338}$).
- Conjectured by Aldous.
- Reason: two independent trees have "different shapes on every scale", so a common subtree would have to "match" large regions of T_n with small regions of T_n'.

The Brownian tree

- Brownian tree $\mathcal{T}$: scaling limit of the trees T_n, with distances renormalized by $\frac{1}{\sqrt{n}}$, and mass $\frac{1}{n}$ on each leaf **[Aldous 90s]**.
- It is a random measured metric space which is compact and has fractal dimension 2.
- Deterministic topology: continuous tree where branching points are dense and have degree 3 **[Croydon–Hambly 07]**.



(picture by I. Kortchemski)

Homeomorphisms of Brownian trees

Theorem (B.–Sénizergues 23+)

Let $\mathcal{T}, \mathcal{T}'$ be two independent Brownian trees. There is $\varepsilon > 0$ such that almost surely, there is no $(1-\varepsilon)$-Hölder homeomorphism from $\mathcal{T}$ to $\mathcal{T}'$.

- Both theorems share most of the proof: partition (R_i) of $\mathcal{T}$ such that for any homeomorphism $\Psi : \mathcal{T} \to \mathcal{T}'$, most of the R_i satisfy $|\Psi(R_i)| \ll |R_i|$.
- To pass from continuous to discrete: classic coupling between $\mathcal{T}$ and T_n (pick n uniform points on $\mathcal{T}$).
- Aldous'proof that $\mathrm{MAST}(T_n, T_n') \geq n^{\frac{\sqrt{3}-1}{2}}$ implicitly builds a Hölder homeomorphism.

THANK YOU !