Common subtrees of independent random trees

(and other common substructures problems)

Caelan Atamanchuk

Department of Mathematics and Statistics
McGill University

CRM-ISM Probability seminar

Acknowledgements

Based on joint work with:

Work initiated at the nineteenth annual Probability and Combinatorics Workshop at the Bellairs Institute in Barbados!

Common substructure problems

1: A TOUR OF COMMON SUBSTRUCTURES.

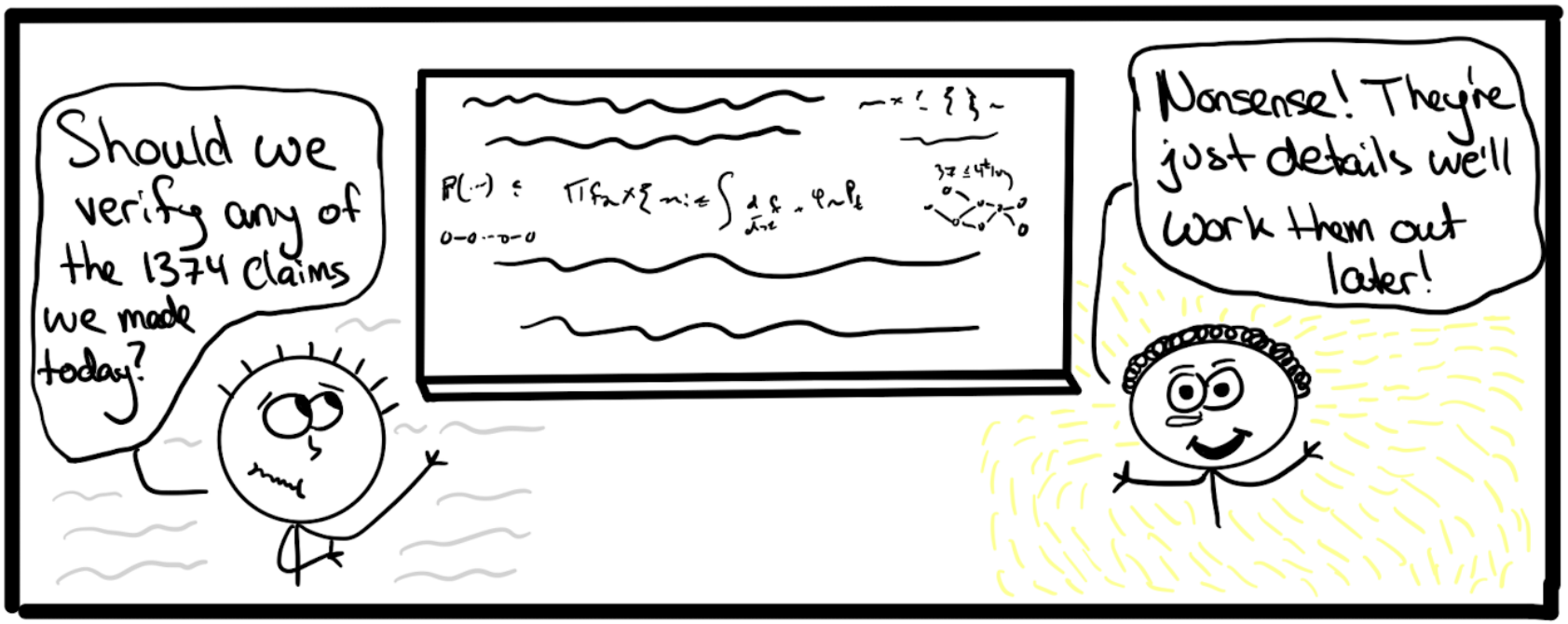
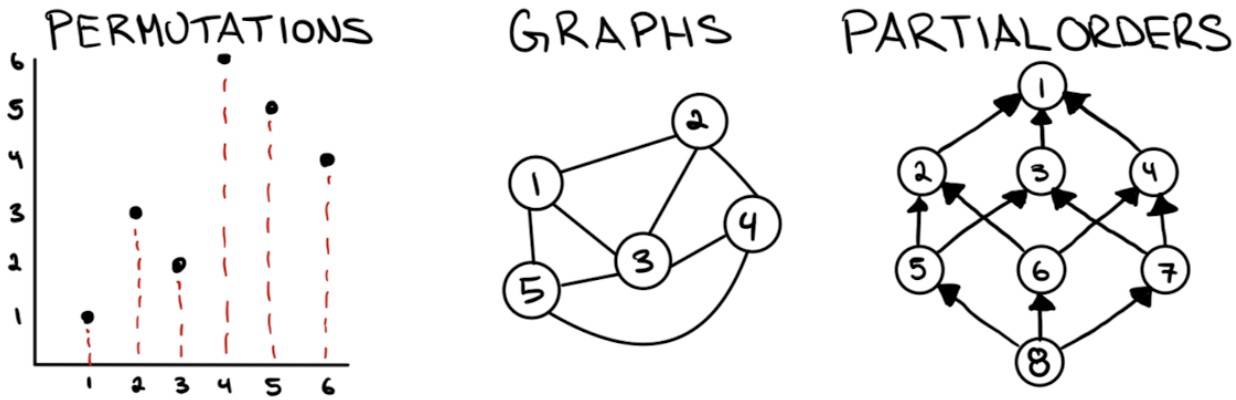


Figure: The two states of mind while doing math on the board.

Common substructure problems

Let $\mathbf{S}_n$ be a collection of combinatorial structures built from $[n]$ with the following property:

- for any $S \in \mathbf{S}_n$ and $A \subseteq [n]$, there is some induced substructure of S on A.



Challenge: For two independent structures $S, S' \in \mathbf{S_n}$ drawn from a probability measure μ_n analyze the following quantity:

$$\left|\operatorname{LCS}(S, S')\right| = \max\left\{|T| : T \text{ a common substructure of } S \text{ and } S'\right\}.$$

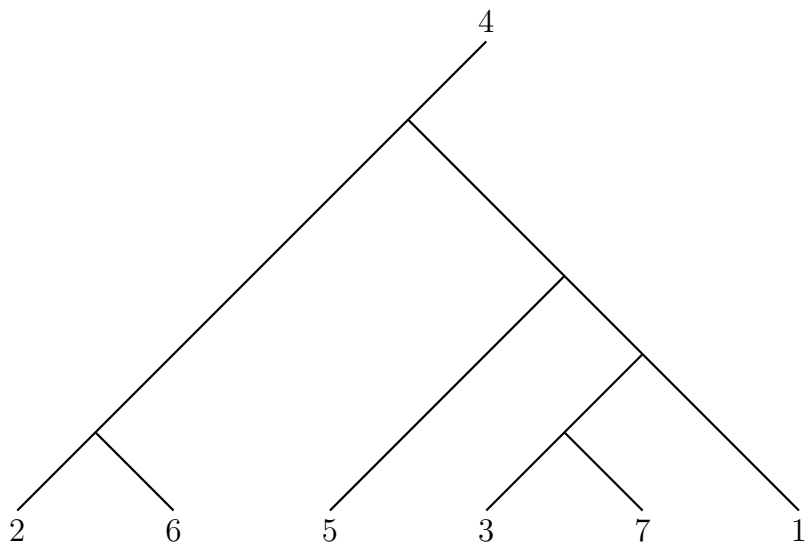
Common substructure problems

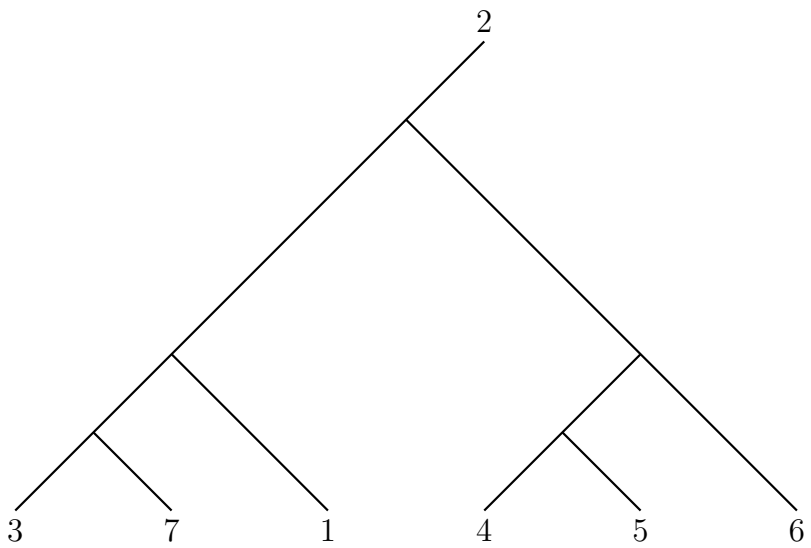
Example: The LCS of two independent strings of length n over the alphabet $\{1, ..., k\}$, X_n and Y_n.

- Many applications in genetics and computational biology.
- By super-additivity, $\mathbf{E}\big[|\operatorname{LCS}(X_n, Y_n)|\big] n^{-1} \to \gamma_k$ as $n \to \infty$. [Chvatal, Sankoff 1975.]
- $\sqrt{k}\gamma_k \to 2$ as $k \to \infty$ [Kiwi, Loebl, Matoušek 2003.].
- $\frac{|\operatorname{LCS}(X_n, Y_n)| - \gamma_k}{\operatorname{Var}(|\operatorname{LCS}(X_n, Y_n)|)} \to N(0, 1)$ as $n \to \infty$. [Houdré, Işlak 2023].

Common substructure problems

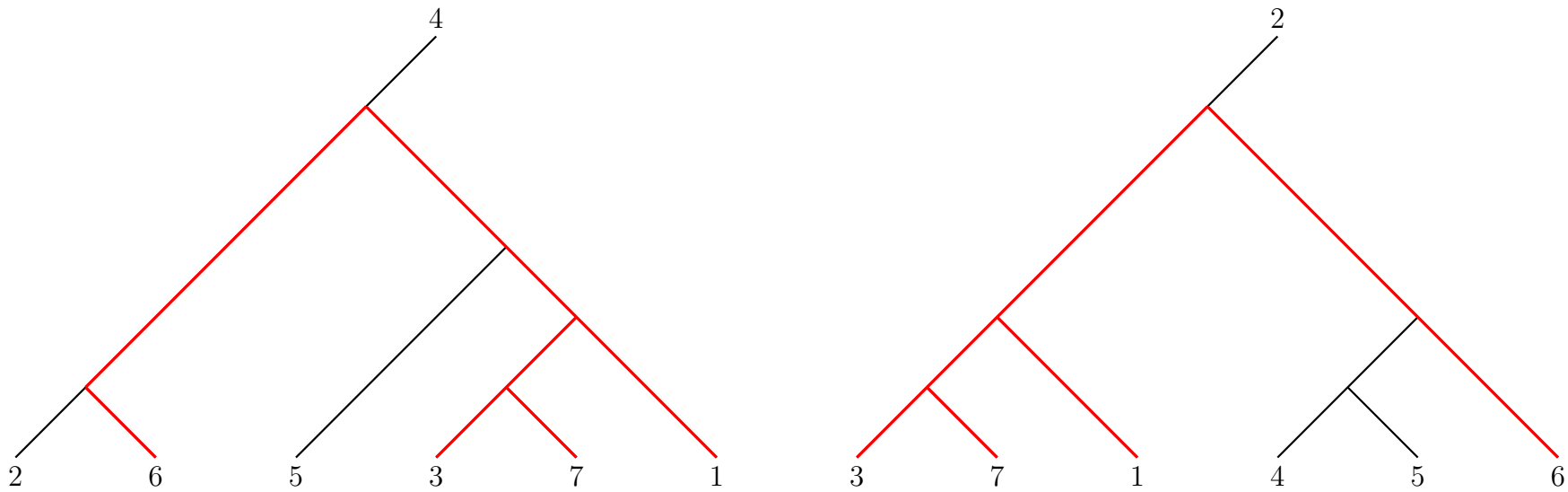
Example: The LCS (or MAST) of two independent uniform cladograms.





Common substructure problems

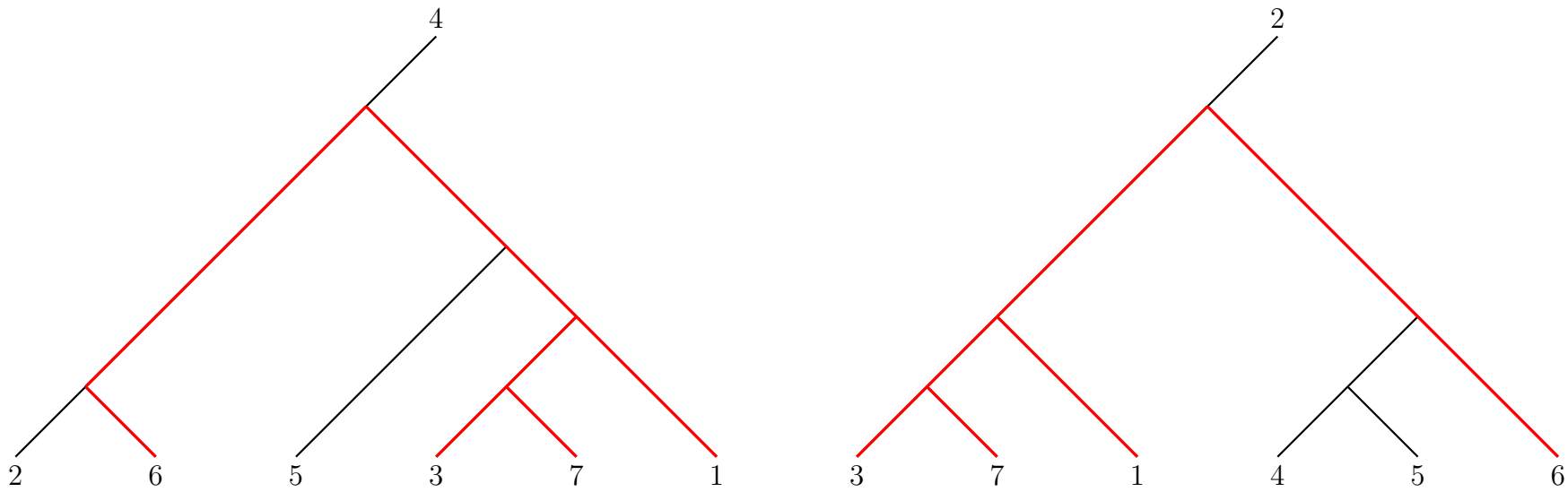
Example: The LCS (or MAST) of two independent uniform cladograms.



- The highlighted parts of the tree is the LCS. The leaves $\{1, 3, 6, 7\}$ have the same ancestral relationships in both trees.

Common substructure problems

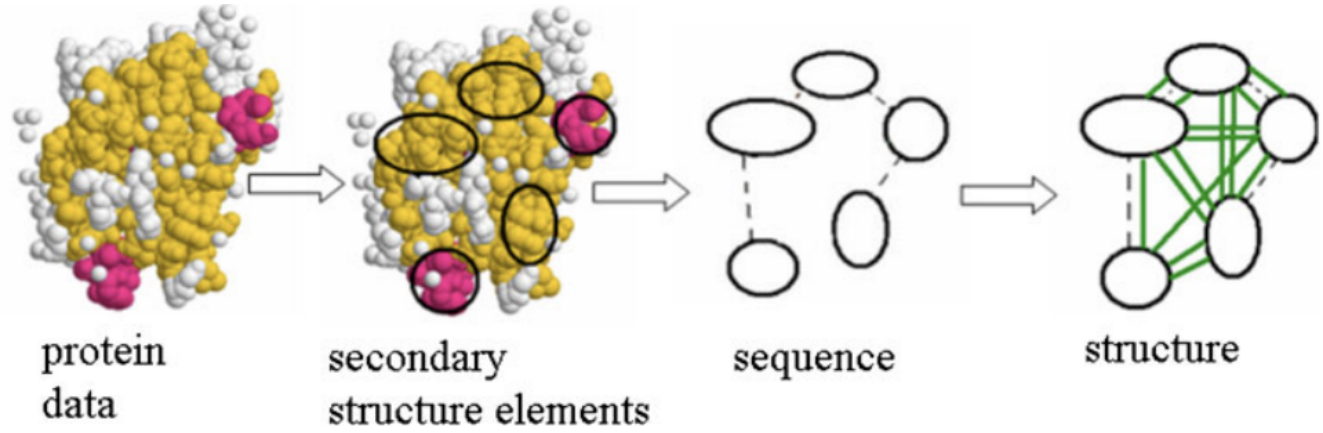
Example: The LCS (or MAST) of two independent uniform cladograms.



Best known bounds: $n^{0.446} \leq \mathrm{LCS} \leq n^{1/2-\epsilon}$.
[Khezeli (2022) and Budzinski, Sénizergues (2023)]
Conjecture: $\mathrm{LCS} = n^{\gamma+o(1)}$ for $\gamma < 1/2$. [Aldous (2022)]

Common substructure problems

Application (network correlation): We can use sizes of common subgraphs as a way to measure similarity in graphs! This has been studied a lot under the name of **graph matching**.

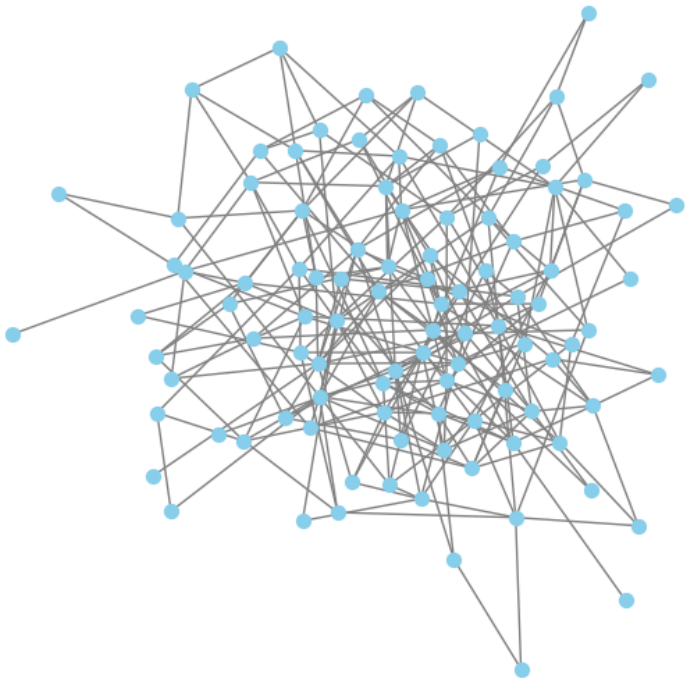


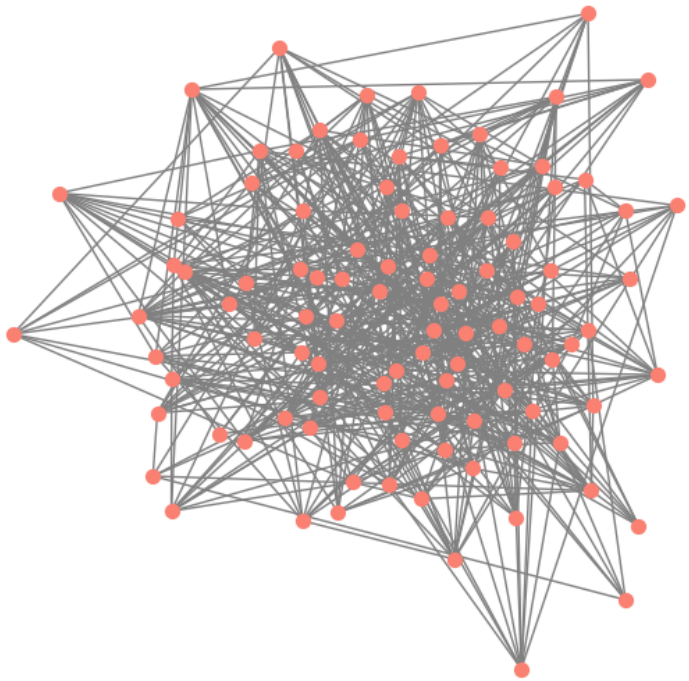
[Livi, Rizzi 2013.]

Common substructure problems

Example: The largest common induced subgraph of two Erdős-Rényi random graphs, with equivalence is up to isomorphism.

- Connects to the famous graph isomorphism problem in computer science.
- The LCS in the dense case has **logarithmic size** and exhibits **two point concentration**. [Chatterjee, Diaconis 2023. Surya, Warnke, Zhu 2025.]





2: The LCS of Bienaymé trees.

Figure: Its easy to forget that everyone is not thrilled by random trees.

Bienaymé trees

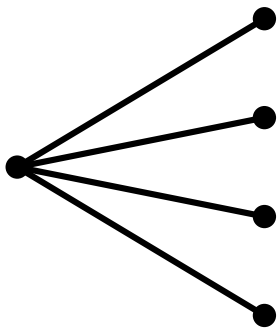
A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.

Bienaymé trees

A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

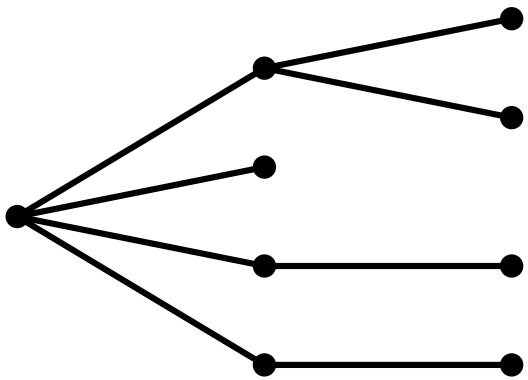
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

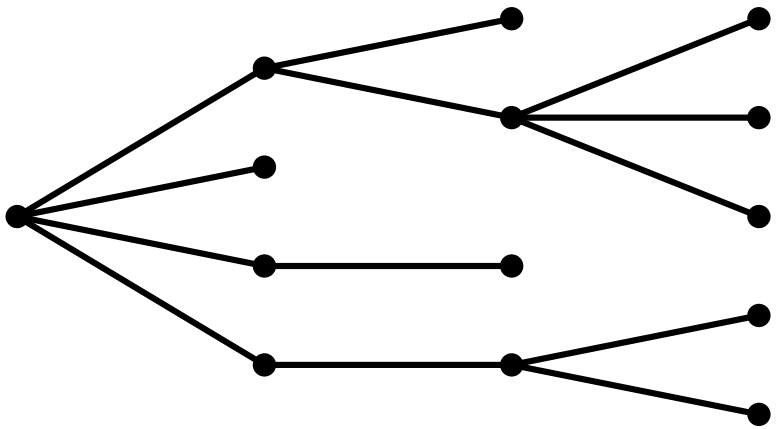
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

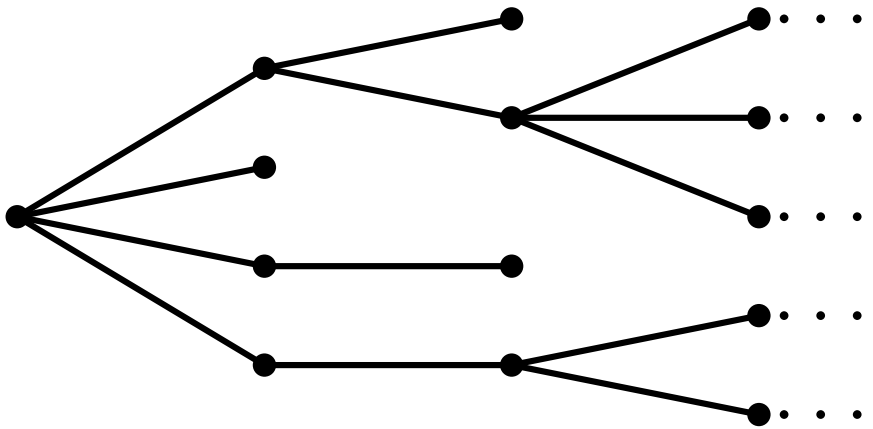
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

Your favourite tree is a Bienaymé tree: By picking specific measures and conditioning on our trees to have size n we get many canonical trees.

- $\mu(d) = 1 - \mu(0) = 1/d$ for some $d \geq 2 \implies$ uniform d-ary tree.
- $\mu = \text{Geometric}(1/2)$ for all $k \geq 0 \implies$ uniform rooted plane tree.
- $\mu = \text{Poisson}(1)$ for all $k \geq 0$ plus a randomly labelling the vertices $\implies$ uniform labelled tree.

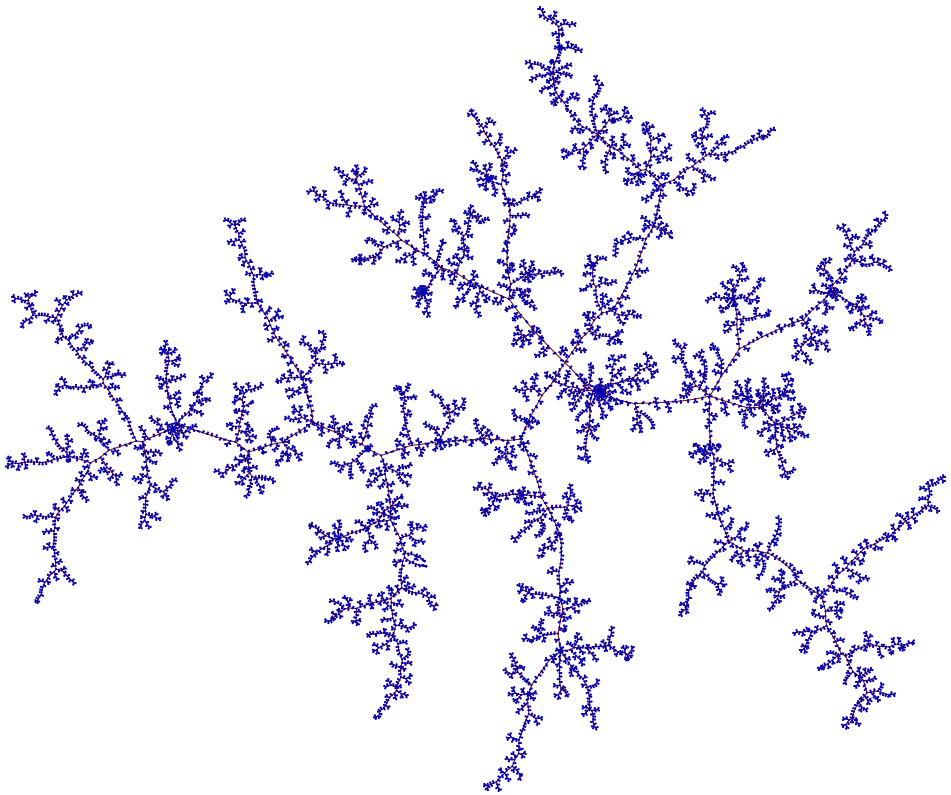
Assumption: μ is such that $\sum_{j=1}^{\infty} j\mu(j) = 1$.
Notation: $\tau_n \leftrightarrow \tau$ conditioned to have size n.

Bienaymé trees

Conditioned Bienaymé trees are quite spiny:

- The height of τ_n is of the order $\sqrt{n}$.
- The distance between uniform random vertices is of order $\sqrt{n}$.

Are common subtrees of them similarly spiny?



[images by Igor Kortchemski!]

LCS of independent Bienaymé trees

Thm (Angel, A., Brandenberger, Donderwinkel, Khanfir 2025+)

Let τ_n and τ_n' be two independent Bienaymé trees conditioned to have size n. Suppose that their offspring distributions, μ and μ', are such that:

- $\sum_{j=1}^{\infty} j^{2+\kappa}\mu(j) < \infty$, and
- $\sum_{j=1}^{\infty} j^{2}\mu'(j) < \infty$.

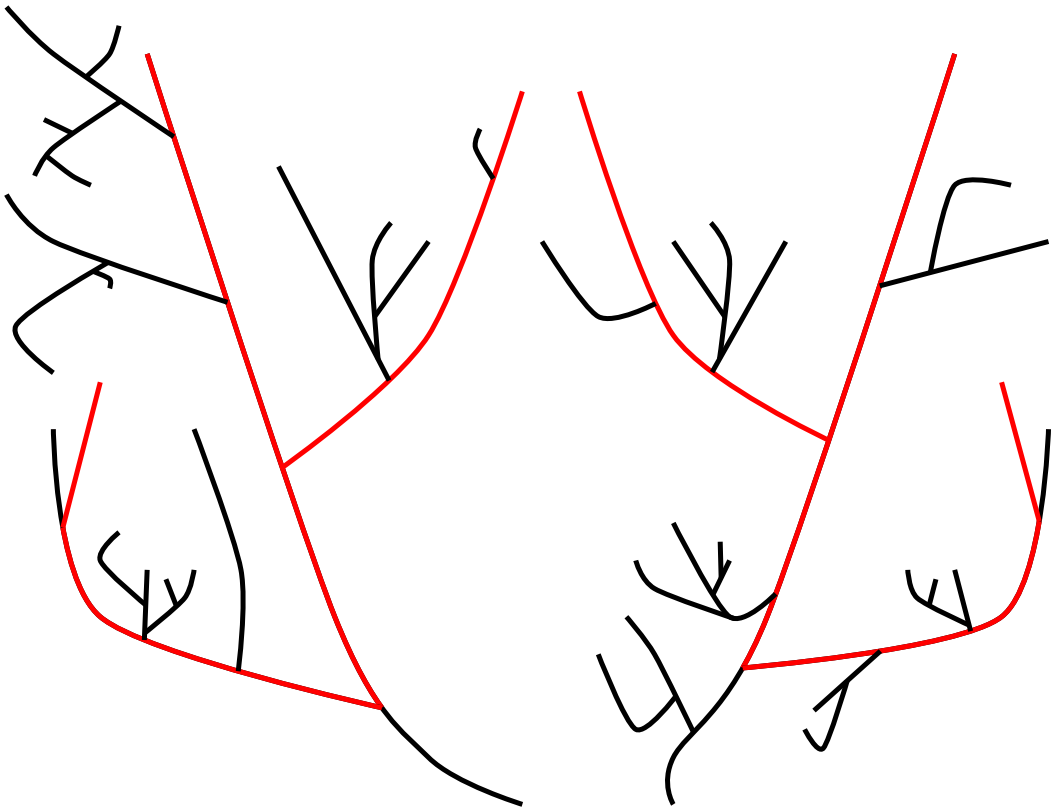
Then, there exists $X > 0$ such that

$$\frac{1}{\sqrt{n}}|\operatorname{LCS}(\tau_n, \tau_n')| \xrightarrow{d} X.$$

TLDR/Heuristic: Large common subtrees under a second and $(2+\kappa)$th moment assumption above are super thin.

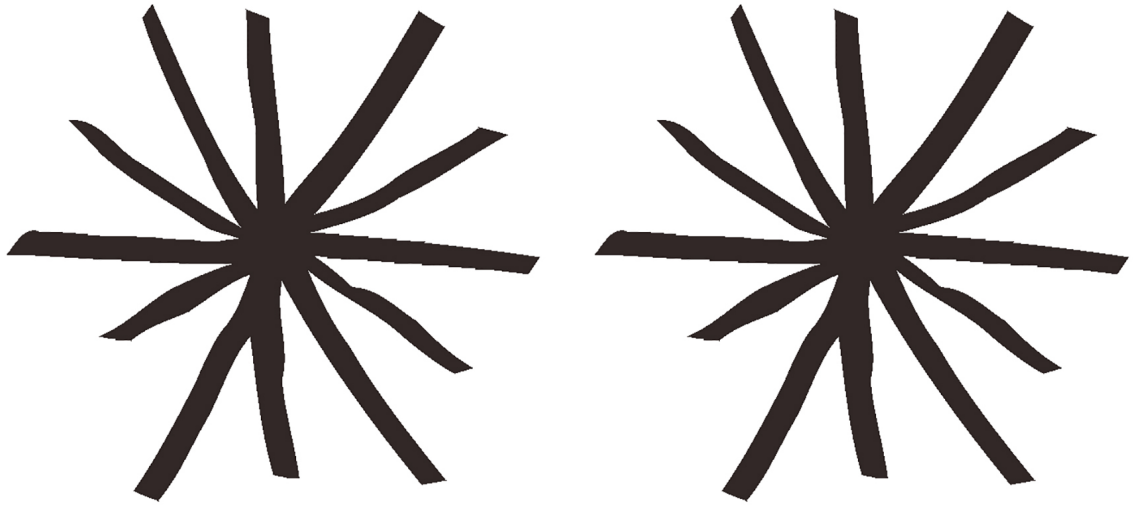
LCS of independent Bienaymé trees

Question: What is the distribution of X?
Answer: The limiting length of the longest common Y between τ_n and τ'_n **up to a constant**.



LCS of independent Bienaymé trees

Question: Is the $2 + \kappa$ moment assumption actually necessary?
Answer: Yes, it is used to avoid large degrees, which allow the creation of common stars...



[Vonnegut 1973]

4: The issue of large degrees.

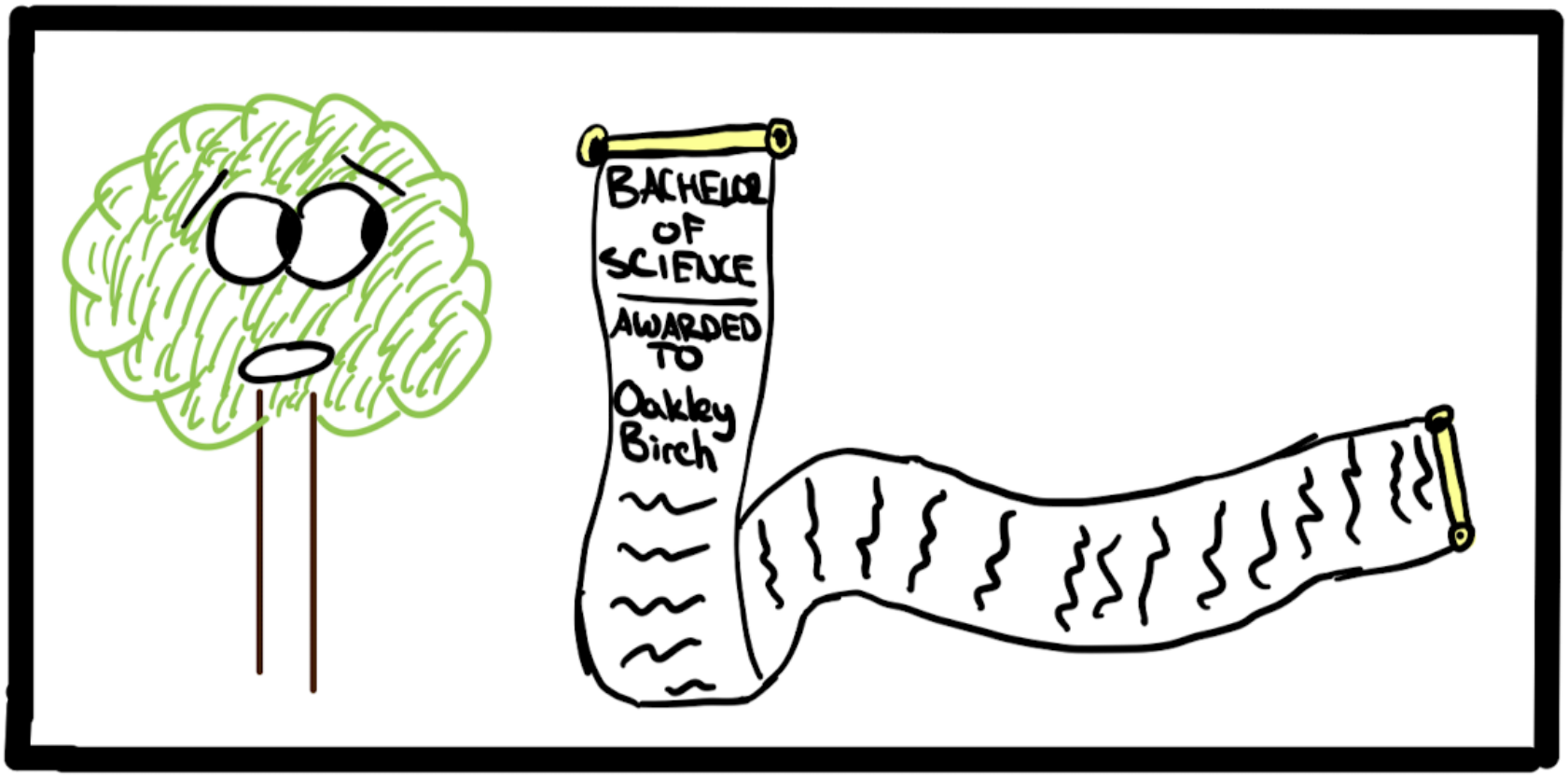
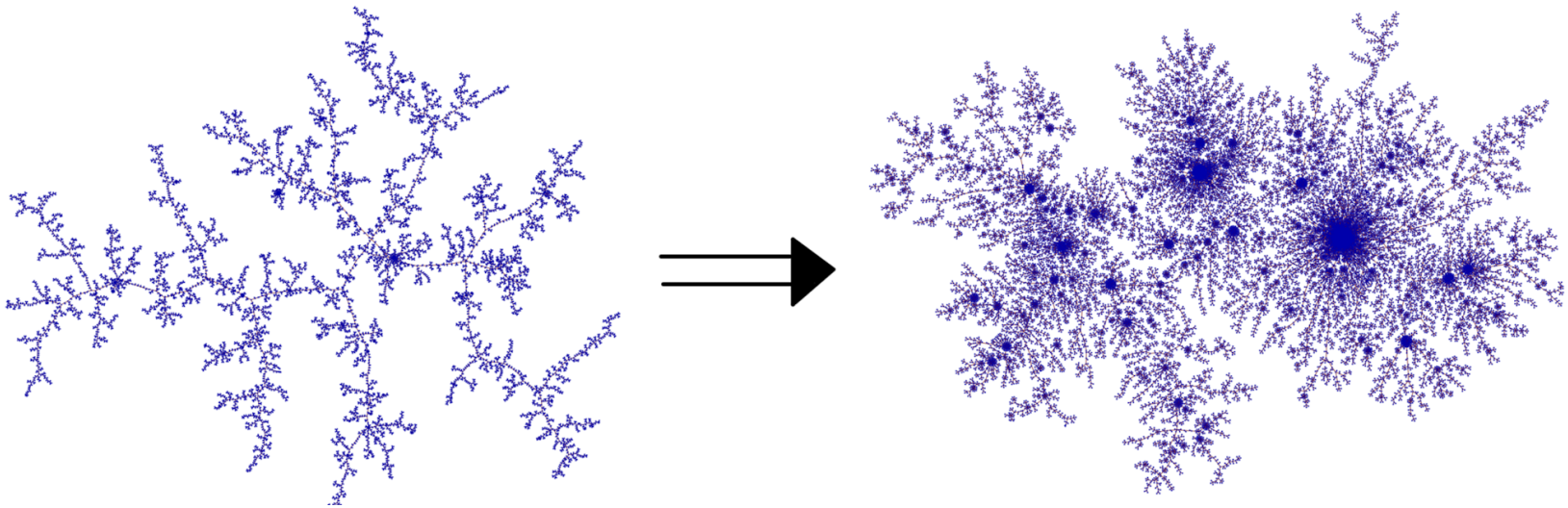


Figure: A tree with a concerningly large degree.

The issue of large degrees

"Facts": Degrees in conditioned Bienaymé trees behave like a sum of i.i.d. random variables. If we assume the largest finite moment of μ is a γth moment for $\gamma > 1$, then we should expect the maximum degree to be order $n^{1/\gamma}$.



- Once the largest degree in our two trees gets too close to $\sqrt{n}$, the LCS should start to change.

The issue of large degrees

Thm (Angel, A., Brandenberger, Donderwinkel, Khanfir 2025+)

Let μ be critical, satisfying $\mu(k) \sim ck^{-3}\log^{-3/2}(k)$. Then, for all $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\liminf_{n\to\infty} \mathbf{P}\Big(\big|\operatorname{LCS}(\tau_n, \tau_n')\big| > \delta \log^{1/4}(n)\sqrt{n}\Big) > 1 - \varepsilon.$$

TLDR: There is a critical offspring distribution such that $\operatorname{LCS}(\tau_n, \tau_n') \geq \log^{1/4}(n)\sqrt{n}$ and

$$\sum_{k\geq 0} \mu(k)k^2 \log^{1/4}(k) < \infty.$$

The issue of large degrees

How to build the counter-example:

- We can find vertices of out-degree $\Theta(\sqrt{n}\log^{-3/4}(n)) = \Delta_n$ in both τ_n and τ'_n.
- the subtrees rooted above a vertex **essentially** behave like independent unconditioned Bienaymé trees.
- The heights of unconditioned Bienaymé trees satisfy $\mathbf{P}(\mathrm{Ht}(\tau) \geq x) \sim cx^{-1}$.

Order the subtrees $\tau_n(1), ..., \tau_n(\Delta_n)$ and $\tau'_n(1), ..., \tau'_n(\Delta_n)$ in decreasing order of height and match the tallest subtrees.

$$\sum_{i=1}^{\Delta_n} \mathrm{Ht}(\tau_n(i)) \wedge \mathrm{Ht}(\tau'_n(i)) \approx \sum_{i=1}^{\Delta_n} \mathrm{Ht}(\tau_n(i)) \approx c\Delta_n \log(\Delta_n).$$

Proof sketch (rooted trees)

3: Why is the LCS thin?

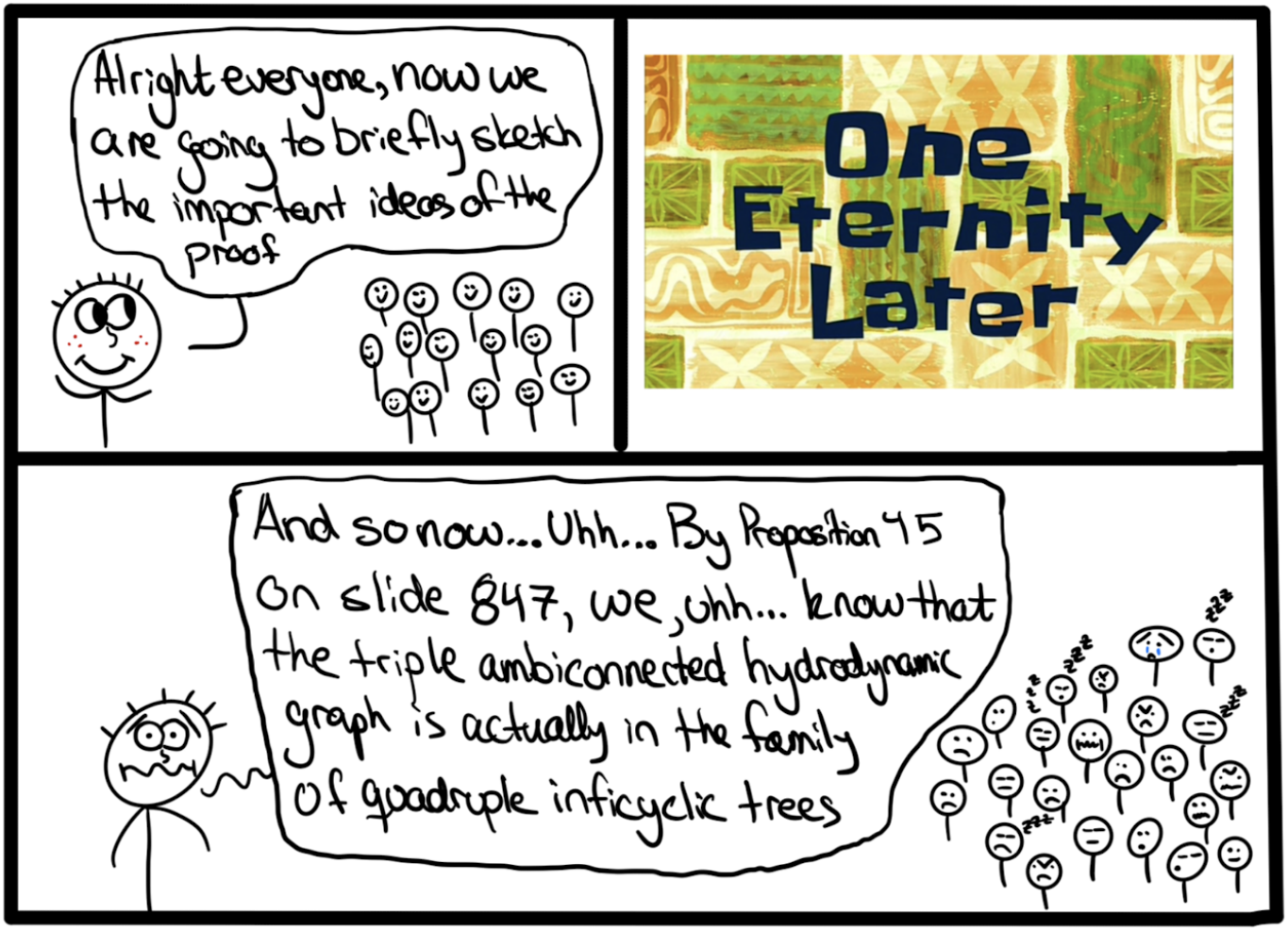


Figure: Trying to explain a proof of your favourite theorem.

Proof sketch (rooted trees)

Theorem

For any $\epsilon > 0$, $\mathbf{P}(|\operatorname{LCS}^{\bullet}(\tau_n, \tau'_n)| \geq n^{1/2+\epsilon}) \to 0$.

Lemma

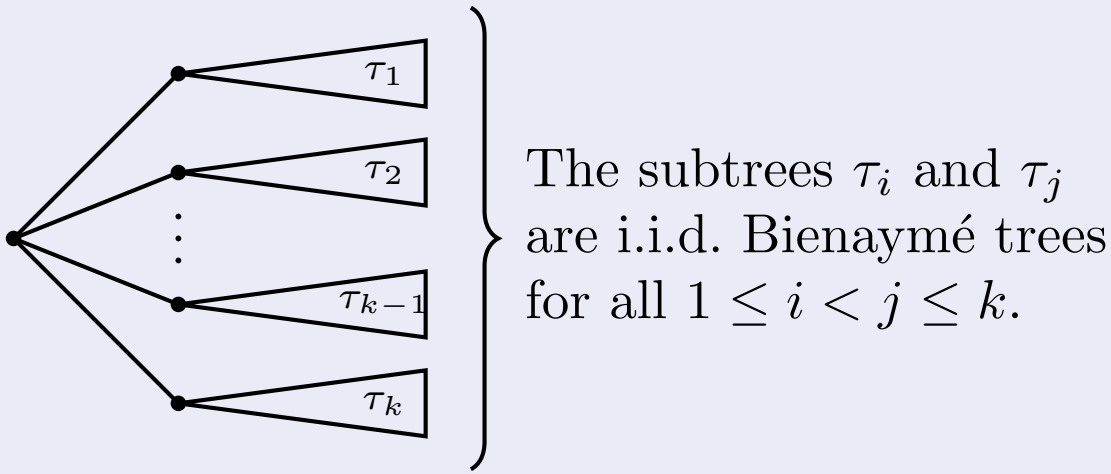
For any $\epsilon, \gamma > 0$ there is a $C > 0$ so that

$$\mathbf{P}\Big(\underbrace{\big\{|\operatorname{LCS}^{\bullet}(\tau, \tau')| \geq h^{1+\epsilon}\big\} \cap \big\{\operatorname{Ht}(\tau) \wedge \operatorname{Ht}(\tau') \leq h\big\}}_{:= P_{\epsilon,h}} \Big) \leq C h^{-\gamma}.$$

Warning: We'll pretend that the two trees are binary.

Proof sketch (rooted trees)

Proposition (the branching property)



Proposition

There exist $c_1, c_2 > 0$ such that $\mathbf{P}(|\tau| = n) \sim c_1 n^{-3/2}$ and $\mathbf{P}(|\tau| \geq n) \sim c_2 n^{-1/2}$.

Proposition (a trivial bound for $|\mathrm{LCS}^\bullet(\tau, \tau')|$)

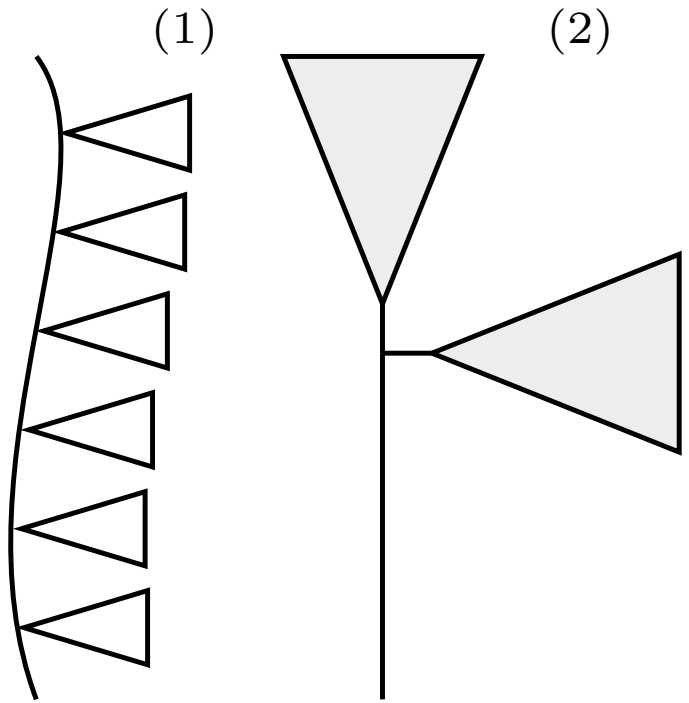
$$\mathbf{P}\big(|\mathrm{LCS}^\bullet(\tau, \tau')| \geq n\big) \leq \mathbf{P}\big(|\tau| \geq n\big)\mathbf{P}\big(|\tau'| \geq n\big) \leq c_2^2 n^{-1}.$$

Proof sketch (rooted trees)

$$P_{\epsilon,h} = \left\{ |\mathrm{LCS}^{\bullet}(\tau, \tau')| \geq h^{1+\epsilon} \right\} \cap \left\{ \mathrm{Ht}(\tau) \wedge \mathrm{Ht}(\tau') \leq h \right\}$$

Idea: Build a path $\mathcal{P}$ in the $\mathrm{LCS}^{\bullet}$ from the root, where we always walk into the largest subtree. There are two cases:

1. Each subtree hanging off of $\mathcal{P}$ is smaller than $h^{1+\epsilon-\nu}$;
2. There is a vertex on $\mathcal{P}$ that has some subtree of size at least $h^{1+\epsilon-\nu}$ hanging off $\mathcal{P}$.

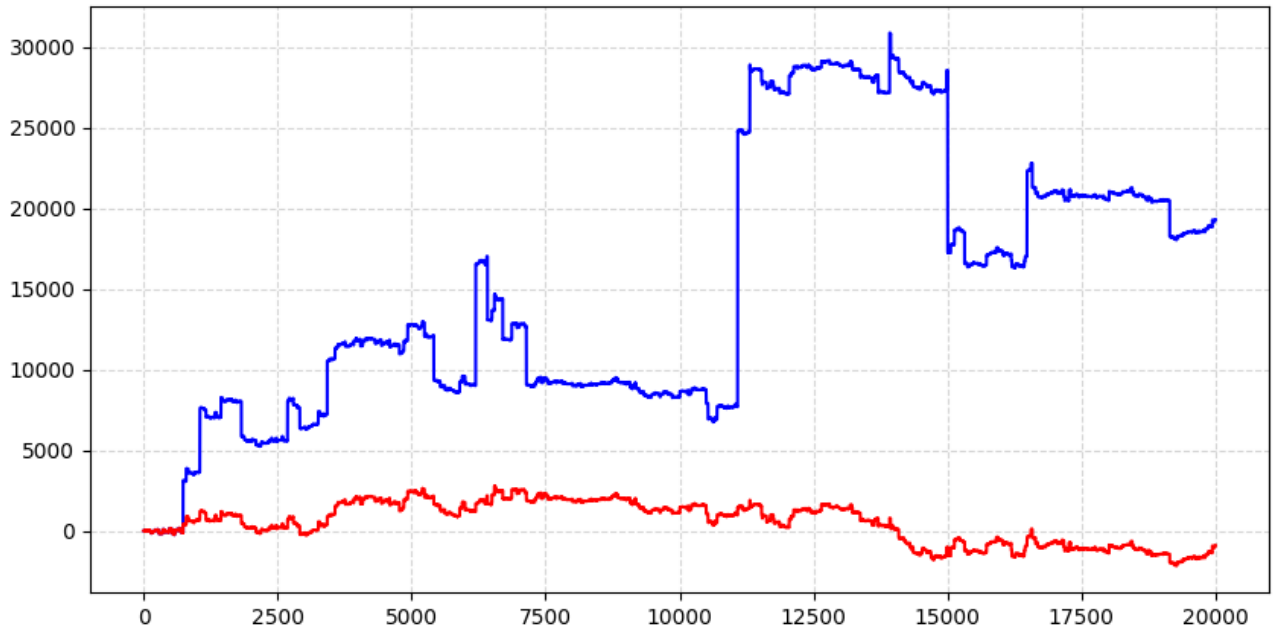


Proof sketch (rooted trees)

One big jump principle (OBJP)

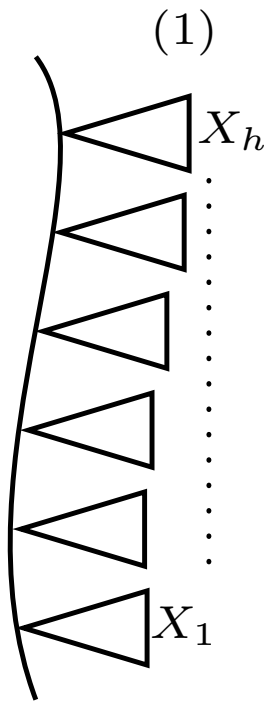
Take $(X_n)_{n=1}^{\infty}$ i.i.d. with $\mathbf{P}(X_i \geq x) \leq cx^{-1}$. For $\gamma > 0$ there exists C such that for any $t, m \geq 0$, $s > 1$, for $S_m = \sum_{i=1}^{m}(X_i \wedge sm^{1+\gamma})$,

$$\mathbf{P}(S_m \geq tm^{1+\gamma}) \leq C\exp(-t/s).$$



Proof sketch (rooted trees)

$$P_{\epsilon,h} = \left\{ |\mathrm{LCS}^\bullet(\tau, \tau')| \geq h^{1+\epsilon} \right\} \cap \left\{ \mathrm{Ht}(\tau) \wedge \mathrm{Ht}(\tau') \leq h \right\}$$



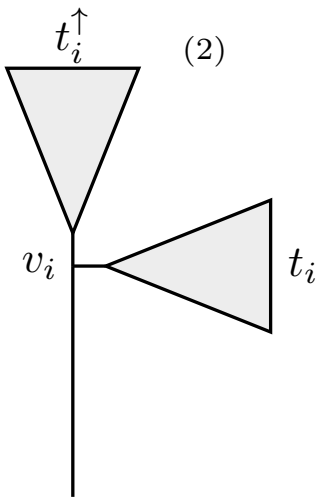
- By branching property and the trivial $\mathrm{LCS}^\bullet$ bound, X_i's are i.i.d. with a distribution that has tails like $\mathbf{P}(X_i \geq x) \leq Cx^{-1}$.
- $|\mathrm{LCS}^\bullet(\tau, \tau')| \leq \sum_{i=1}^{h} (X_i \wedge h^{1+\epsilon-\nu})$ by definition of $P_{\epsilon,h}$ and (1).
- We can apply the OBJP with $t = h^{\nu/2}$, $s = 1$, and $\gamma = \nu/2$!

Conclusion:

$$\mathbf{P}\Big(P_{\epsilon,h} \cap (1)\Big) \leq Ch^2 \exp\left(-h^{\nu/2}\right)..$$

Proof sketch (rooted trees)

$$P_{\epsilon,h} = \left\{ |\mathrm{LCS}^{\bullet}(\tau, \tau')| \geq h^{1+\epsilon} \right\} \cap \left\{ \mathrm{Ht}(\tau) \wedge \mathrm{Ht}(\tau') \leq h \right\}$$



By construction of $\mathcal{P}$, $|t_i| \geq h^{1+\epsilon-\nu}$ and $|t_i^{\uparrow}| \geq h^{1+\epsilon-\nu}$.
We can use a union bound and the trivial LCS bound:

$$\begin{aligned}
\mathbf{P}(P_{\epsilon,h} \cap (2)) &\leq Ch\mathbf{P}(P_{\epsilon-\nu,h})^2 \\
&\leq Ch\mathbf{P}(P_{\epsilon-\nu,h})\mathbf{P}(|t_i| \geq h^{1+\epsilon-\nu}) \\
&\leq C\mathbf{P}(P_{\epsilon-\nu,h})\frac{1}{h^{\epsilon-\nu}}.
\end{aligned}$$

Conclusion [cases (1) and (2)]:

$$\mathbf{P}(P_{\epsilon,h}) \leq C\mathbf{P}(P_{\epsilon-\nu,h})\frac{1}{h^{\epsilon-\nu}} + Ch^2 \exp\left(-h^{\nu/2}\right)$$

Proof sketch (rooted trees)

Repeatedly applying the inequality from the previous slide gives:

Proposition

For all $\epsilon, \gamma > 0$ there is a $C > 0$ such that $\mathbf{P}(P_{\epsilon,h}) \leq Ch^{-\gamma}$.

Corollary

For any $\epsilon > 0$, $\mathbf{P}(|\operatorname{LCS}^{\bullet}(\tau_n, \tau'_n)| \geq n^{1/2+\epsilon}) \to 0$.

proposition $\implies$ corollary:

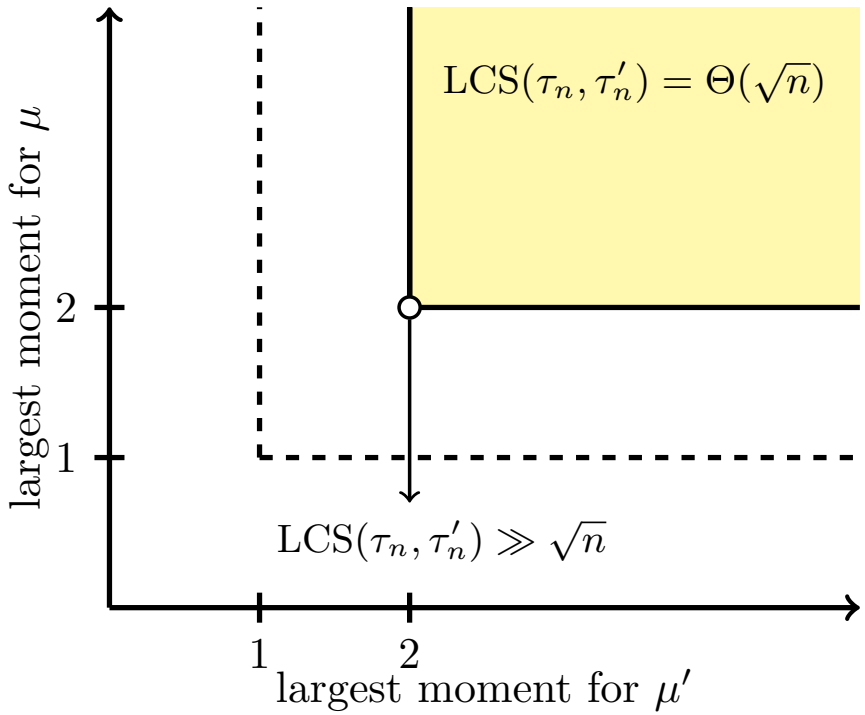
- Choose γ large enough that $\mathbf{P}\big(P_{\epsilon/2, n^{1/2+\epsilon/2}} \mid |\tau| = |\tau'| = n\big) \to 0$.
- Apply previous results that imply $\mathbf{P}\big(\operatorname{Ht}(\tau) \wedge \operatorname{Ht}(\tau') \geq n^{1/2+\epsilon/2} \mid |\tau| = |\tau'| = n\big) \to 0$.

5: Cool things for the future.

Figure: My application to NSERC for funding (rejected).

Future directions

- What if the two trees are not the same size? For example, take τ_n and τ'_m, where $m = n^\alpha$ for some $\alpha \in (0, 1)$.
- What happens if we allow some distortion or sample the trees with dependence?
- Other moment assumptions?
- and much much more...



Future directions

- Thank you all for listening! These slides, as well as a mostly comprehensive list of common substructure references, are available on my website.

↓↓ QR code for the references :) ↓↓