Common subtrees of independent random trees
(and other common substructures problems)

Caelan Atamanchuk

Department of Mathematics and Statistics
McGill University

University of Victoria Probability and Dynamics seminar

Acknowledgements

Based on joint work with:

Robin Khanfir

Anna Brandenberger

Omer Angel

Serte Donderwinkel

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

1: A TOUR OF COMMON SUBSTRUCTURES.

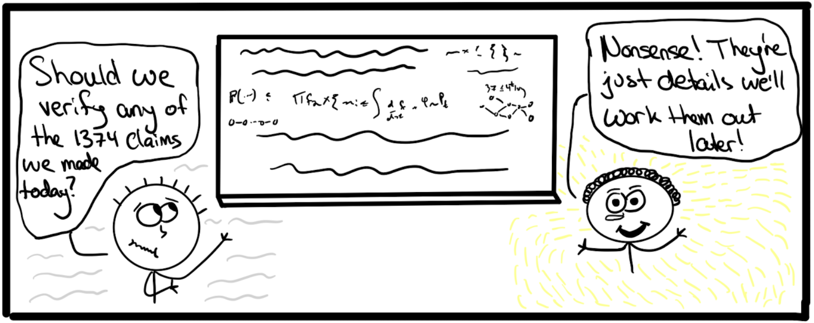
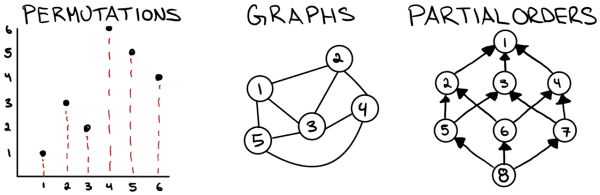


Figure: The two states of mind while doing math on the board.

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:

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

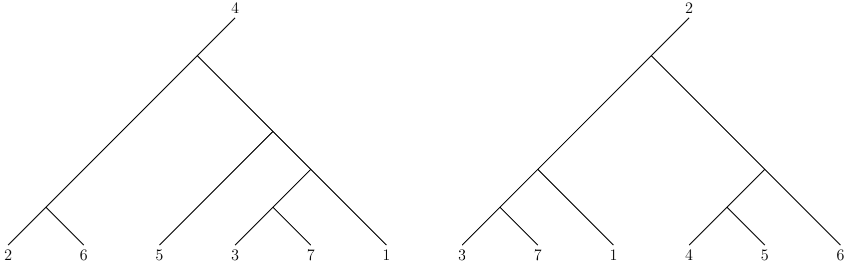
Common substructure problems

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, $\mathbf{E}\left[|\operatorname{LCS}(X_n, Y_n)|\right] n^{-1} \to \gamma_k$ as $n \to \infty$. [Chvatal, Sankoff 1975.]
- $\sqrt{k}\gamma_k \to 2$ as $k \to \infty$ [Kiwi, Loebl, Matoušek 2003.].
- $\frac{|\operatorname{LCS}(X_n, Y_n)| - \gamma_k}{\mathrm{Var}(|\operatorname{LCS}(X_n, Y_n)|)} \to N(0, 1)$ as $n \to \infty$. [Houdré, Işlak 2023].

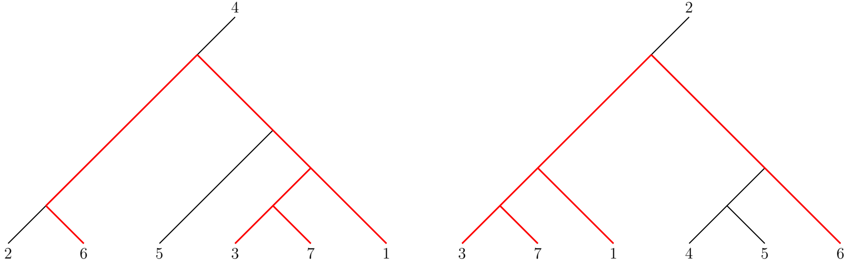
Common substructure problems

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



Common substructure problems

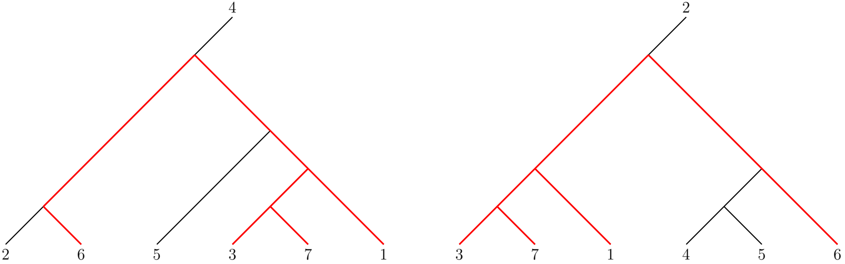
Example: The LCS (or MAST) of two independent uniform cladograms.



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

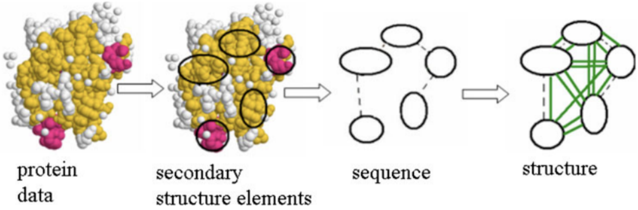
Common substructure problems

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



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

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**.

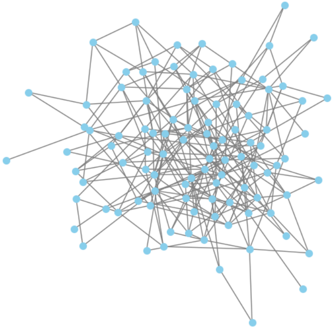


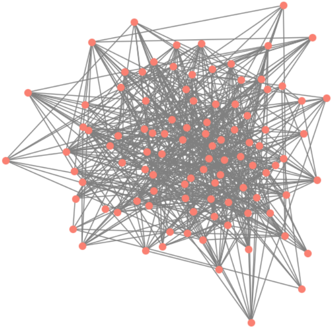
[Livi, Rizzi 2013.]

Common substructure problems

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

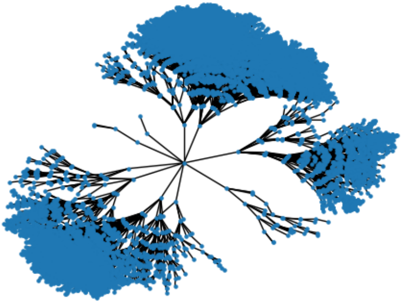
- Connects to the famous graph isomorphism problem in computer science.
- The LCS in the dense case has **logarithmic size** and exhibits **two point concentration**. [Chatterjee, Diaconis 2023. Surya, Warnke, Zhu 2025.]





Common substructure problems

Example: The largest common subtree of two independent uniform random recursive trees. [Baumler, Kerriou, Martin, Lodewijks, Powierski, Rácz, Sridhar 2025+.]



Best bounds: $n^{0.83} \leq \mathrm{LCS}(T_n, T_n') \leq 0.99n$
Conjecture: $\mathrm{LCS}(T_n, T_n') = n^{1-o(1)}$.

Bienaymé trees

2: The LCS of conditioned Bienaymé trees.

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

Bienaymé trees

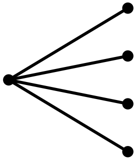
A random tree model: Given μ a measure on $\{0, 1, 2, ...\}$ we define a random plane tree T as follows:

- 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:

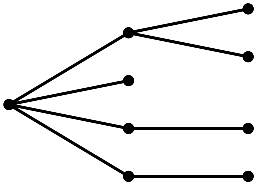
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

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

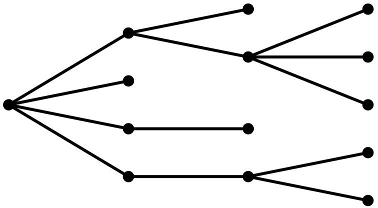
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



Bienaymé trees

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

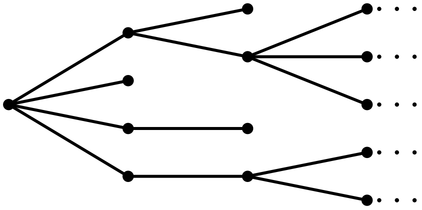
- Start with a root, it has a random number of children drawn from μ.
- Given the tree up to generation k, give each vertex in generation k a number of children drawn independently from μ.



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

- We will focus on trees with offspring distribution μ such that $\sum_{j=1}^{\infty} j\mu(j) = 1$ and $\sigma^2 = \sum_{j=0}^{\infty} \mu(j)(j-1)^2 < \infty$. These are right on the boundary of being finite.
- Specifically, we care about large–n asymptotics of critical Bienaymé trees conditioned to have size n.

Proposition [Kesten, Ney, and Spitzer 1966]

$\mathbf{P}(\mathrm{Ht}(\tau) \geq x) \sim \frac{2}{x\sigma^2}$.

Proposition [Folklore 1900s]

$\mathbf{P}(|\tau| = n) \sim c_1 n^{-3/2}$ and $\mathbf{P}(|\tau| \geq n) \sim c_2 n^{-1/2}$.

Bienaymé trees

Your favourite tree is a Bienaymé tree: By picking μ carefully 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.

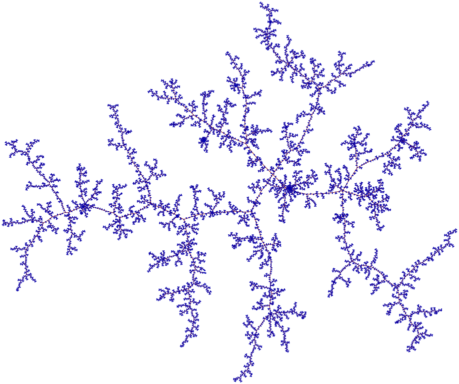
Notation: $\tau_n \leftrightarrow \tau$ conditioned to have size n.

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 with a mutual offspring distributions μ such that

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

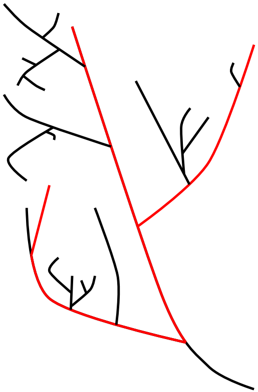
Then, there exists $X > 0$ such that

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

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

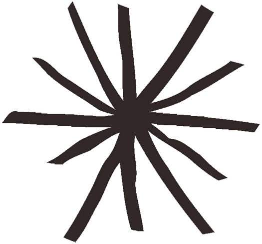
LCS of independent Bienaymé trees

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 depending on** μ. (A Y is a subtree with exactly one degree 3 vertex, and no degree ≥ 4 vertex.)



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. There are counterexamples when the $2+\kappa$ condition fails.



[Vonnegut 1973]

4: The issue of large degrees.

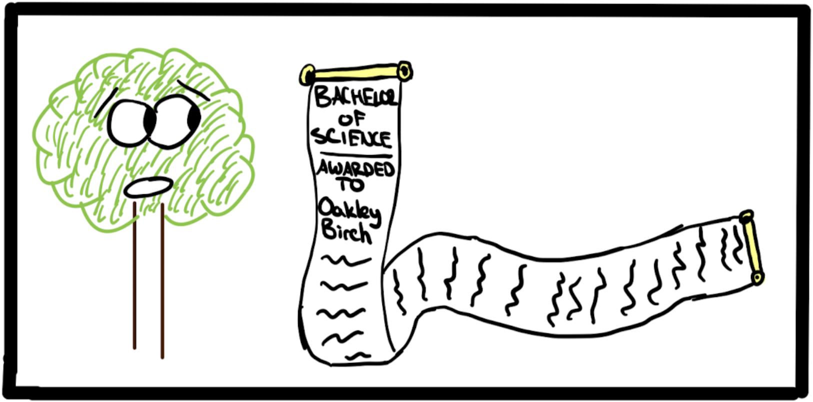
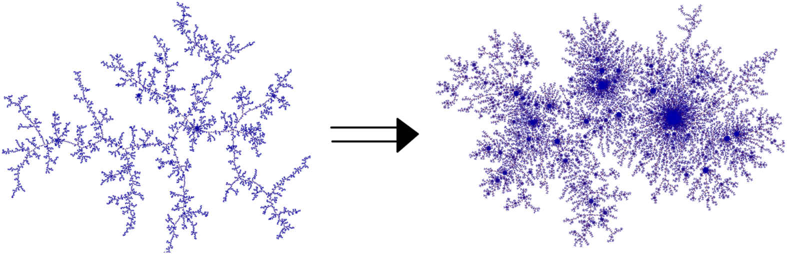


Figure: A tree with a concerningly large degree.

The issue of large degrees

"Facts": Degrees in conditioned Bienaymé trees behave like i.i.d. μ distributed 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 starts 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

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

In particular, there is a critical offspring distribution such that $\mathrm{LCS}(\tau_n, \tau_n') \geq \log^{1/4}(n)\sqrt{n}$ and

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

The issue of large degrees

Extreme value theory: Let $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ be non-negative i.i.d. random variables with tails like $\mathbf{P}(X_1 \geq x) = \mathbf{P}(Y_1 \geq x) \sim cx^{-1}$.

The ith order statistic of $(X_i)_{i=1}^n$ and $(Y_i)_{i=1}^n$ (the ith largest entry of the respective vectors), $X^{(i)}$ and $Y^{(i)}$ are both close in order of magnitude to $\frac{n}{i}$. Thus,

$$\sum_{i=1}^{n} (X^{(i)} \wedge Y^{(i)}) \asymp \sum_{i=1}^{n} \frac{n}{i} \wedge \frac{n}{i} \asymp n \log(n)$$

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.

$$|\mathrm{LCS}(\tau_n, \tau_n)| \geq \sum_{i=1}^{\Delta_n} \mathrm{Ht}(\tau_n(i)) \wedge \mathrm{Ht}(\tau_n'(i)) \asymp \Delta_n \log(\Delta_n).$$

3: Why is the LCS thin?

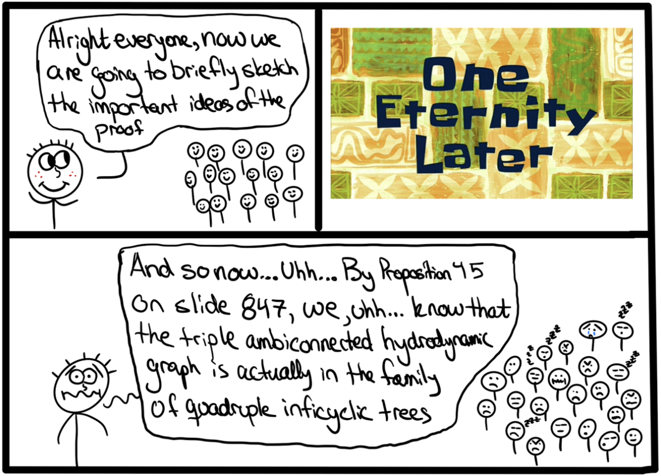


Figure: Trying to explain a proof of your favourite theorem.

Proof sketch (rooted trees)

Theorem

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

Lemma

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

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

Lemma

For any $\epsilon, \nu > 0$, $\mathbf{P}(P_{\epsilon,h}) \leq C\mathbf{P}\big(P_{\epsilon-\nu,h}\big)\frac{1}{h^{\epsilon-\nu}} + Ch^2 \exp\big(-h^{\nu/2}\big)$.

Proof sketch (rooted trees)

Theorem

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

Lemma

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

Proving Theorem using Lemma:

- From last slide we can choose γ large enough that

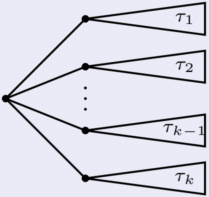
$$\mathbf{P}\left(P_{\epsilon, n^{1/2+\epsilon}} \mid |\tau| = |\tau'| = n\right) \to 0.$$

- From known results about Bienaymé tree heights we know that

$$\mathbf{P}\left(\mathrm{Ht}(\tau) \wedge \mathrm{Ht}(\tau') \geq n^{1/2+\epsilon} \mid |\tau| = |\tau'| = n\right) \to 0.$$

Proof sketch (rooted trees)

Proposition (the branching property)



The subtrees τ_i and τ_j are i.i.d. Bienaymé trees for all $1 \leq i < j \leq k$.

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 linear bound for $|\operatorname{LCS}^\bullet(\tau, \tau')|$)

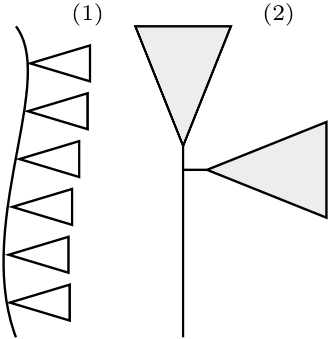
$$\mathbf{P}(|\operatorname{LCS}^\bullet(\tau, \tau')| \geq n) \leq \mathbf{P}\big(|\tau| \geq n\big)\mathbf{P}\big(|\tau'| \geq n\big) \sim c_2^2 n^{-1}.$$

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 $\mathcal{P}$.

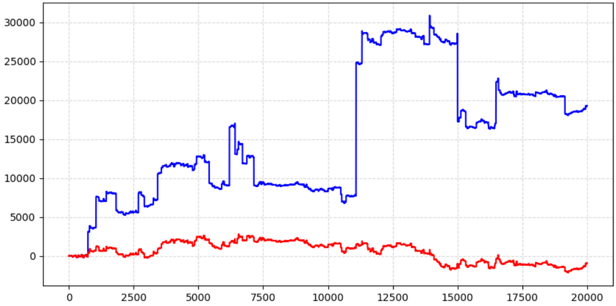


Proof sketch (rooted trees)

Proposition (tails for truncated sums)

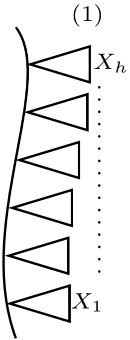
Take $(X_n)_{n=1}^{\infty}$ i.i.d. with $\mathbf{P}(X_i \geq x) \leq cx^{-1}$. For $\gamma > 0$ there exists C such that for any $t, m \geq 0$, $s > 1$, for $S_m = \sum_{i=1}^{m}(X_i \wedge sm^{1+\gamma})$,

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



Proof sketch (rooted trees)

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



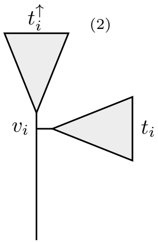
- By branching property and the linear $\operatorname{LCS}^\bullet$ bound, X_i's are i.i.d. with a distribution that has tails like $\mathbf{P}(X_i \geq x) \leq Cx^{-1}$.
- $|\operatorname{LCS}^\bullet(\tau,\tau')| \leq \sum_{i=1}^{h}(X_i \wedge h^{1+\epsilon-\nu})$ by definition of $P_{\epsilon,h}$ and (1).
- We can apply the tail bounds from the last slide 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 linear 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)]:

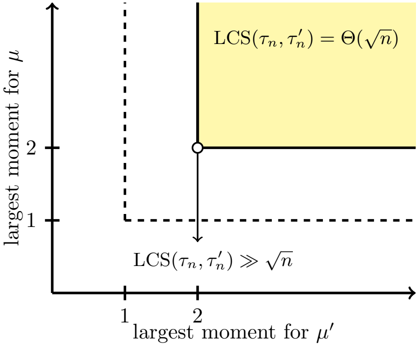
$$\mathbf{P}(P_{\epsilon,h}) \leq C\mathbf{P}(P_{\epsilon-\nu,h})\frac{1}{h^{\epsilon-\nu}} + Ch^2 \exp\left(-h^{\nu/2}\right)$$

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.

$\downarrow\downarrow$ QR code for the references :) $\downarrow\downarrow$