Sparsification of Phylogenetic Covariance Matrices of k-Regular Trees

Sean Svihla

Department of Applied Mathematics
University of Colorado - Boulder

AofA 2024

k-Regular Trees

$T = (V, E)$:

- **planted**
- **ordered**
- **unlabelled**
- k-**regular**, $k \geq 2$.

Leaves labelled as encountered in DFS.

Interior nodes and **edges** labelled by leaf descendants.

Edge length function
$\ell : E \longrightarrow \mathbb{R}_+$.

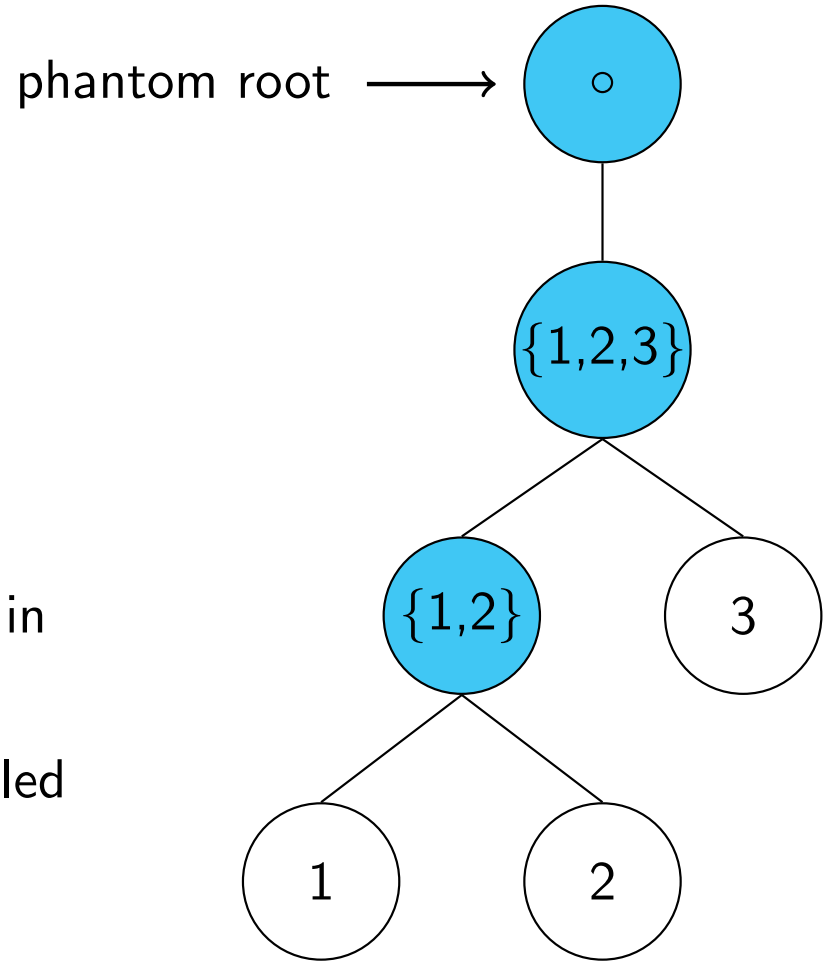


Figure: 2-regular tree example.

Phylogenetic Covariance Matrices

(a.k.a. Cophenetic Matrices)

Phylogenetic covariance is a function of shared path length:

$$C(i, v) = \ell(e_1);$$
$$C(i, j) = \ell(e_2) + \ell(e_1);$$
$$C(i, i) = \ell(e_3) + \ell(e_2) + \ell(e_1).$$

Phylogenetic covariance matrix:

$$C = \Big(C(i,j)\Big)_{i,j \in L}.$$

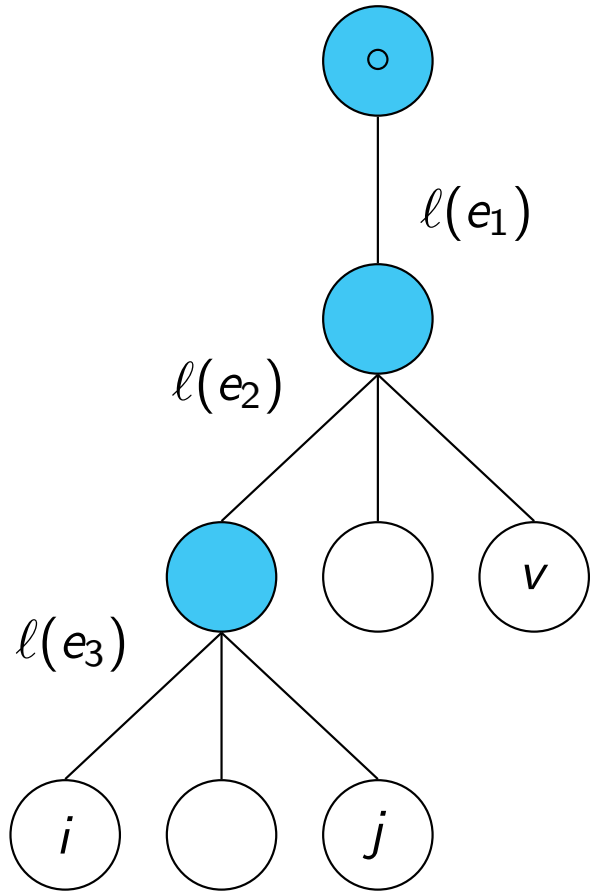


Figure: 3-regular tree with interior nodes colored cyan.

Phylogenetic Covariance

[Harmon (2019)]

If traits evolve as a **Brownian Motion** (BM) along each edge, i.e.:

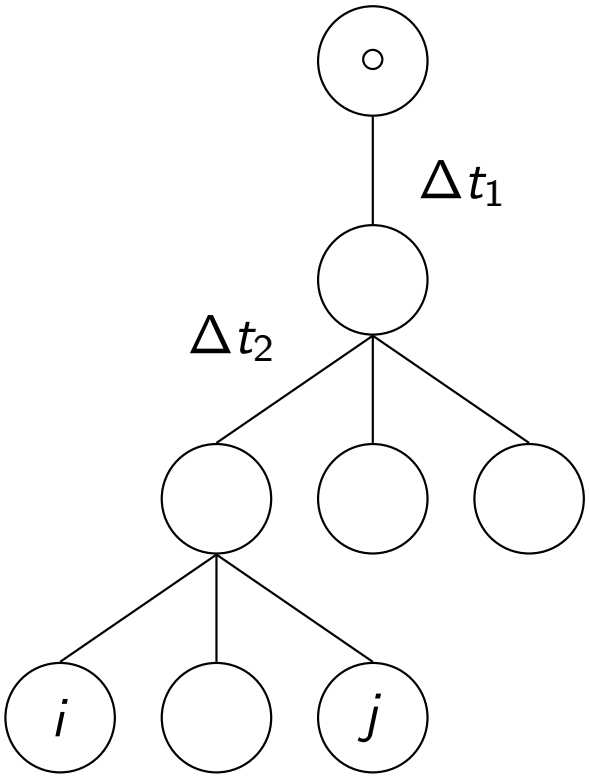


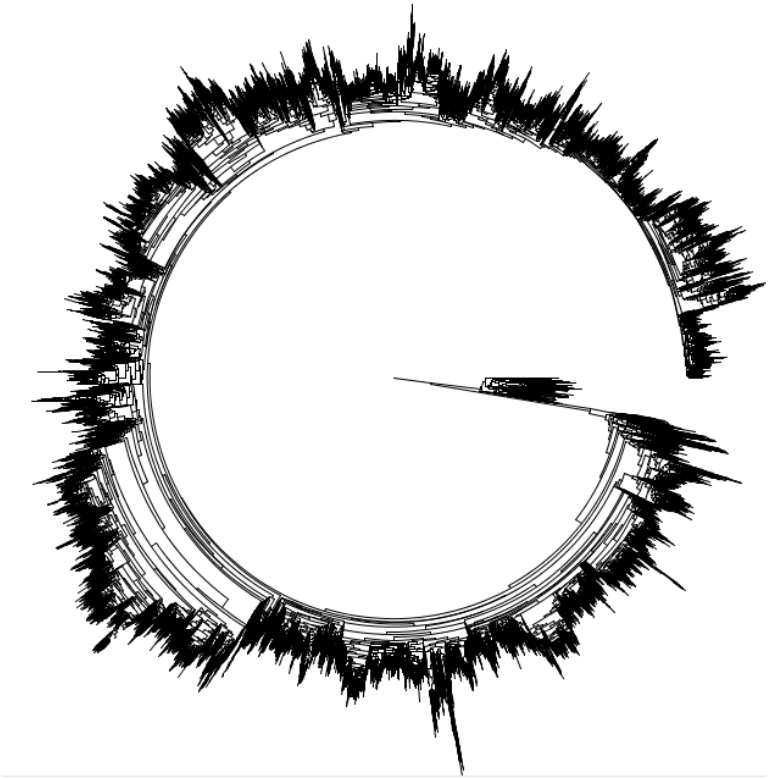
Figure: BM representation of phylogenetic covariance matrices.

$$X_v := \text{trait value at a leaf } v$$
$$U, V := \text{Gaussian variables}$$

then U, V, and $B(\Delta t_1 + \Delta t_2)$ are independent, hence

$$\begin{aligned} X_i &= B(\Delta t_1 + \Delta t_2) + U \\ X_j &= B(\Delta t_1 + \Delta t_2) + V \\ &\Downarrow \\ \mathrm{Cov}(X_i, X_j) &= \mathrm{Var}\big(B(\Delta t_1 + \Delta t_2)\big) \\ &= (\Delta t_1 + \Delta t_2) \cdot \sigma^2 \\ &= \ell(e_1) + \ell(e_2). \end{aligned}$$

Phylogenetic covariance matrices are typically dense



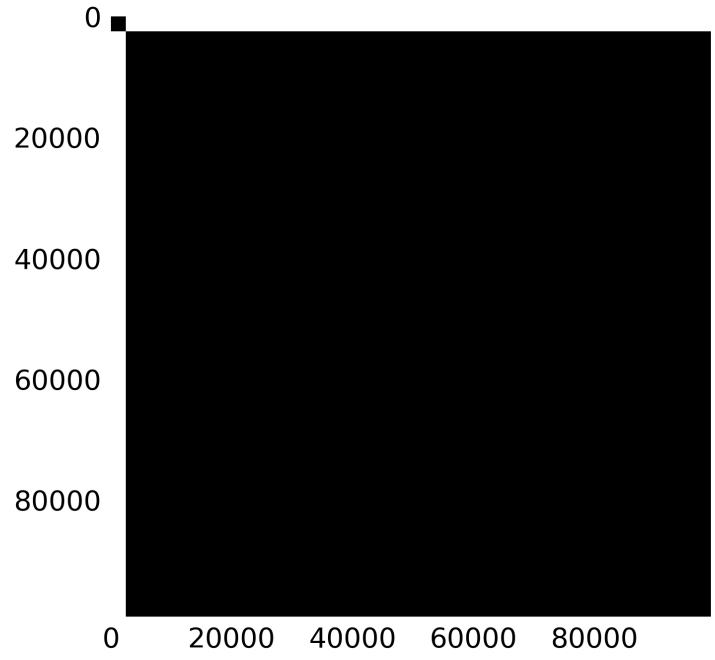
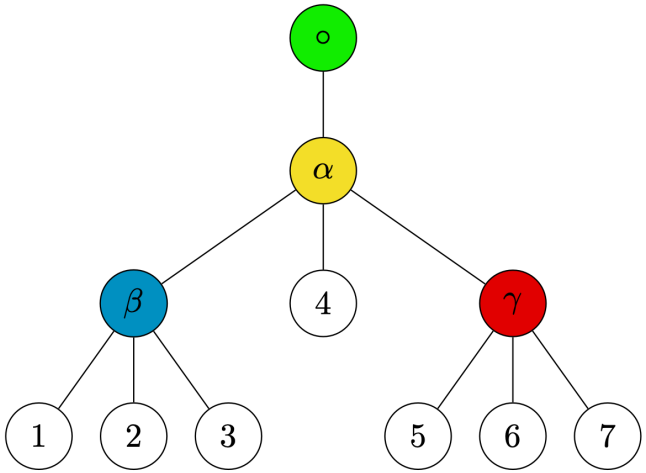


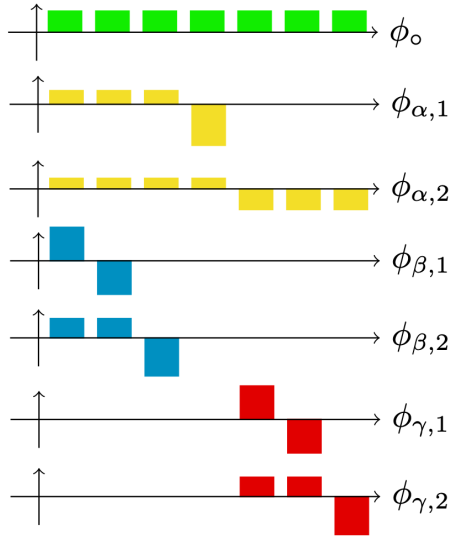
Figure: **(Left)** Circular layout of reference binary phylogenetic tree with ≈100,000 leaves. **(Right)** Heatmap of associated covariance matrix. 94% of its ≈10 billion entries are non-zero.[a]

[a] Figures from [Gorman & Lladser (2023)].

The Haar-like Wavelets

[Gavish, Nadler & Coifman (2010)]





- **Orthonormal basis** for the linear space of functions $L \to \mathbb{R}$.
- As many wavelets as leaves.
- The wavelet associated with the root is **constant**.

For each $v \in I \setminus \{\circ\}$

- there are associated wavelets $\phi_{v,j}$, $j = 1, 2, \ldots, |\text{children}(v)| - 1$,
- $\forall j$, $\text{supp}(\phi_{v,j}) \subset L(v)$, and
- $\forall j$, $\phi_{v,j}$ takes a single positive and single negative value.

The Haar-like Wavelets

[Gavish, Nadler & Coifman (2010)]

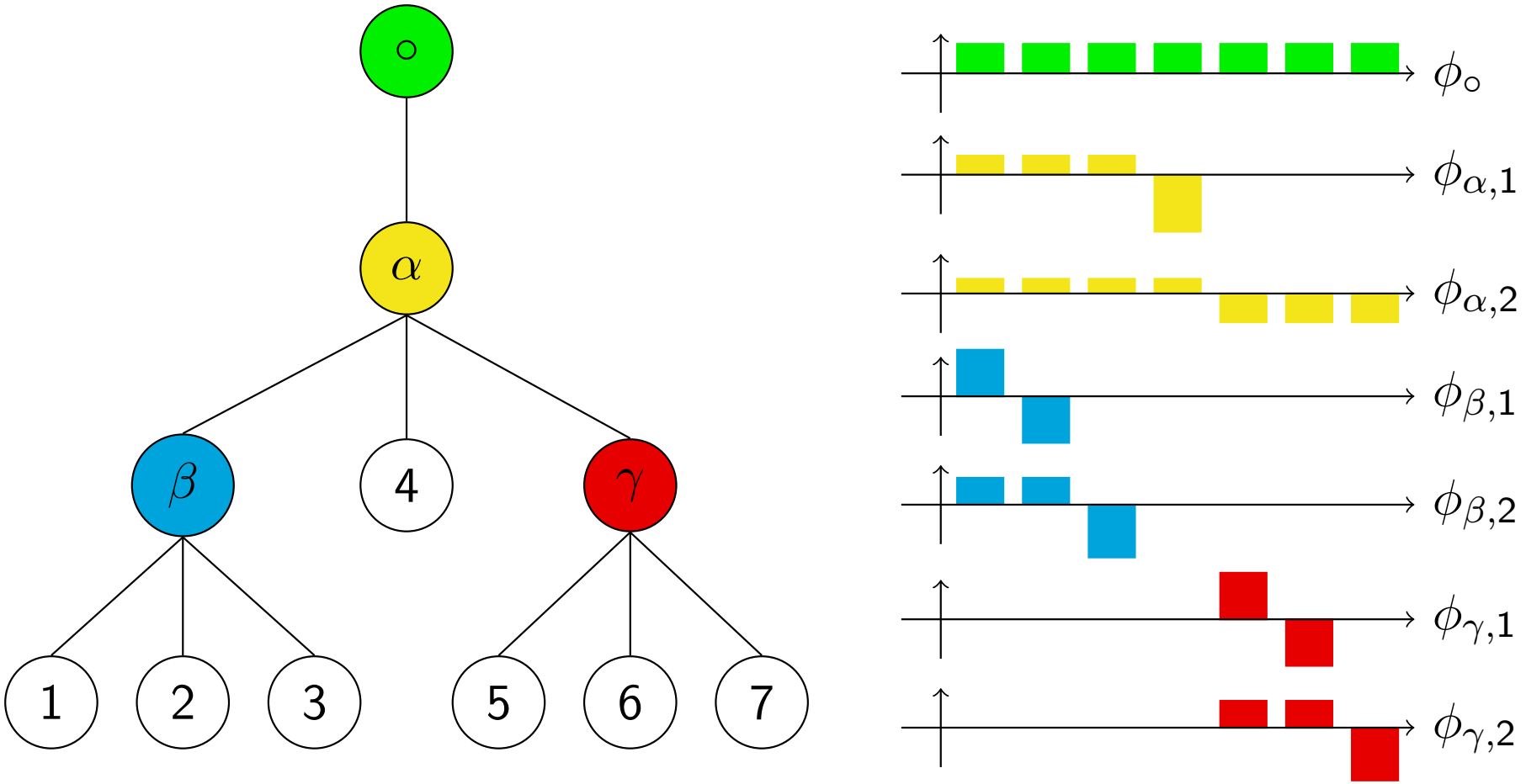


Figure: Haar-like wavelets associated with a 3-regular tree.

Sparsification of phylogenetic covariance matrices

Trace branch length is defined $\forall e \in E,\ \ell^*(e) := |L(e)| \cdot \ell(e)$.

Theorem.[a] Let Φ be the matrix with Haar-like wavelets as columns. If $u, v \in I$ then

$$\big(\Phi' C \Phi\big)(u, v) = \sum_{i \in L} \varphi_u(i) \cdot \ell^*(i, v) \cdot \varphi_v(i).$$

In particular, $L(u) \cap L(v) = \emptyset \implies \big(\Phi' C \Phi\big)(u, v) = 0$.

[a] [Gorman & Lladser (2023)], [Svihla & Lladser (2024)].

Many wavelet pairs have disjoint support

Recall: $\forall \varphi_u$ wavelet associated with $u \in I$, $\mathrm{supp}(\varphi_u) \subset L(u)$.

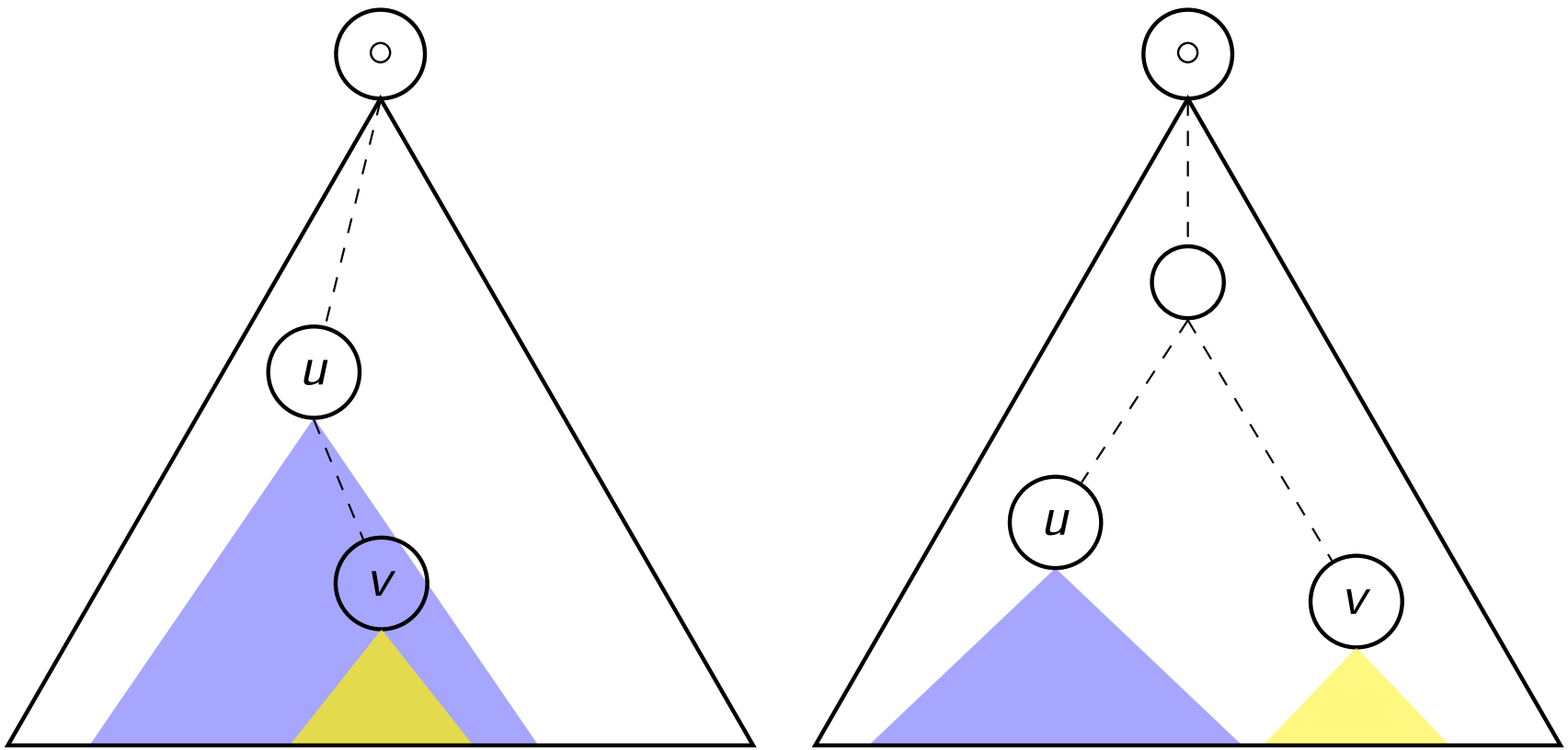
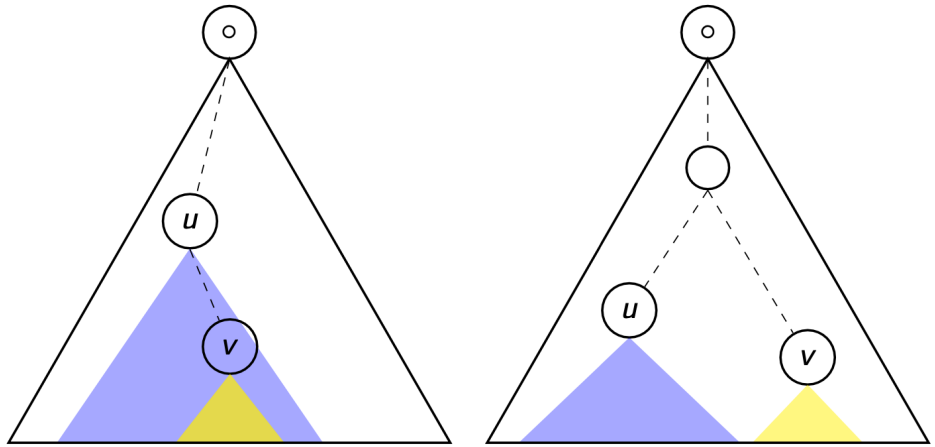


Figure: $L(u) \cap L(v) \neq \emptyset \iff$ "u is an ancestor of v or vice versa".

Sparsification of phylogenetic covariance matrices



Theorem.[b] For a k-regular tree T, if ζ denotes the fraction of vanishing entries of $\Phi' C \Phi$, then

$$(1-\zeta) \leq \frac{(k-1)^2}{|I|} + 2(k-1)^2 \frac{\mathrm{IPL}(T)}{|I|^2}.$$

In particular, if $\mathrm{IPL}(T) \ll |I|^2$ then $\zeta = 1 - \mathrm{o}(1)$.

[b] [Svihla & Lladser (2024)].

There's no reason why $\mathrm{IPL}(T) \ll |I|^2$ should happen!

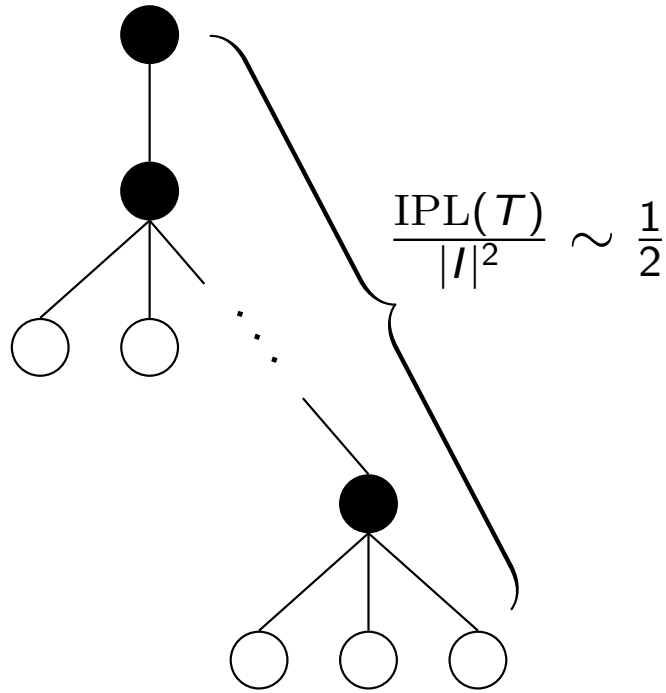


Figure: 3-regular caterpillar tree.

What about a typical k-regular tree?

Generating function for k-regular trees

Let $\mathbb{T}_n$ be a **uniformly at random k-regular tree** of size n.

Goal: Find asymptotic formulas for $\mathbb{E}\big[\mathrm{IPL}(\mathbb{T}_n)\big]$ and $\mathbb{V}\big[\mathrm{IPL}(\mathbb{T}_n)\big]$.

Definition. Let $Q(z,u)$ be the (bivariate) generating function of the class of k-regular trees, where z marks the **size** and u marks the **internal path length** of each tree.

To address the goal, we need to understand the singularities of

$$\mathrm{Q}(\mathrm{z}) := \mathrm{Q}(\mathrm{z},\mathrm{u})\Big|_{u=1};$$

$$\mathrm{Q}_u(z) := \frac{\partial Q}{\partial u}(z,u)\Big|_{u=1}; \quad Q_{uu}(z) := \frac{\partial^2 Q}{\partial u^2}(z,u)\Big|_{u=1}.$$

Radius of convergence of $Q(z)$

By the **symbolic method**:

$$Q(z) = 1 + z\{Q(z)\}^k$$

By **Lagrange inversion**:

$$[z^n]Q(z) = \frac{1}{(k-1)n+1}\binom{kn}{(k-1)n}$$

So, $Q(z)$ has **radius of convergence** and a **singularity** at:

$$z_k := \frac{(k-1)^{k-1}}{k^k}.$$

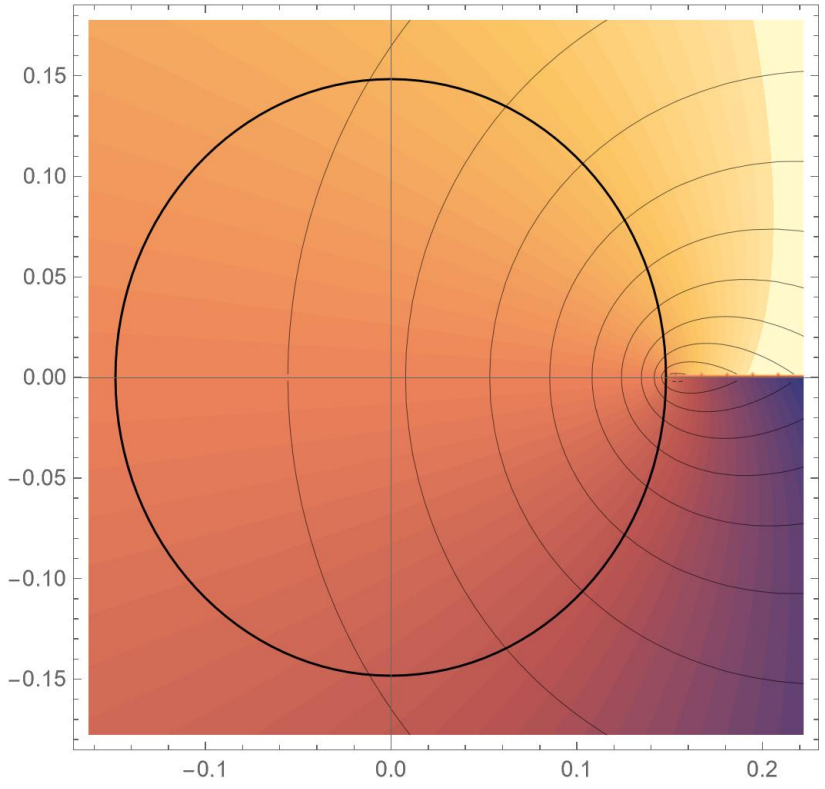


Figure: Plot of $Q(z)$ and boundary of disk of convergence $|z| < z_k$ when $k = 3$. Plot is colored by argument; contour lines denote modulus.

Interlude: Hypergeometric functions

Definition. $F(z) = \sum_{n=0}^{\infty} f_n z^n$ is called **hypergeometric** if there are $a_1, \ldots, a_p, b_1, \ldots, b_q \in \mathbb{R}$ such that

$$\frac{f_{n+1}}{f_n} = \frac{(n+a_1)\cdots(n+a_p)}{(n+1)\cdot(n+b_1)\cdots(n+b_q)}.$$

In this case, we write:

$$F(z) = {}_pF_q\left[\begin{matrix} a_1, \ldots, a_p \\ b_1, \ldots, b_q \end{matrix}; z\right].$$

Theorem[†] If the **balance** $s := \left(\sum_{j=1}^{q} b_j - \sum_{j=1}^{p} a_j\right) > 0$ then ${}_pF_q(a_1, \ldots, a_p; b_1, \ldots, b_q; z)$ converges at $z = 1$.

[†] [Evans & Stanton (1984)].

$Q(z)$ is hypergeometric

Proposition.[c]

$$Q(z) = {}_{k-1}F_{k-2}\left[\begin{matrix}\frac{1}{k}, \frac{2}{k}, \ldots, \frac{k-1}{k} \\ \frac{2}{k-1}, \frac{3}{k-1}, \ldots, \frac{k-2}{k-1}, \frac{k}{k-1}\end{matrix}; \frac{z}{z_k}\right]$$

and it has balance $s > 0$. Furthermore, if we define

$$p(t) := t(1-t)^{k-1},$$

then

$$Q(p(t)) = \frac{1}{1-t}, \text{ for all } 0 \leq t \leq \frac{1}{k}.$$

[c][Weisstein (2023)], [Svihla & Lladser (2024)].

In particular, since $z_k = p\left(\frac{1}{k}\right)$, $Q(z_k) = \frac{k}{k-1}$.

$Q(z)$ fits the smooth implicit function schema

$$Q(z_k) = \frac{k}{k-1}$$

Lemma.[d] $z = z_k$ is the only singularity of $Q(z)$ on $|z| \leq z_k$, and

$$Q(z) = 1 + g(z) - h(z) \cdot \sqrt{1 - \frac{z}{z_k}},$$

locally about z_k, with $g(z)$ and $h(z)$ analytic nearby. Additionally,

$$[z^n]\, Q(z) = \sqrt{\frac{k}{2\pi n^3 (k-1)^3}} \cdot z_k^{-n} \left(1 + \mathrm{O}\left(n^{-1}\right)\right).$$

[d] [Drmota (2009)], [Svihla & Lladser (2024)].

Partial derivatives of $Q(z,u)$ at $u=1$

Again due to the **symbolic method**:

$$Q(z,u) = 1 + z \cdot \left\{ Q(zu,u) \right\}^k.$$

Implicit differentiation then gives that $Q_u(z)$ and $Q_{uu}(z)$ are **linear combinations** of generating functions of the form

$$\frac{f(z)\{Q(z)\}^a}{(1-kz\{Q(z)\}^{k-1})^b}.$$

Lemma.[e] The equation $kz\left\{Q(z)\right\}^{k-1} = 1$, with $|z| \leq z_k$, has only z_k as a solution.

[e][Svihla & Lladser (2024)].

Partial derivatives of $Q(z, u)$ at $u = 1$

Lemma.[f] If $f : \mathbb{C} \to \mathbb{C}$ is an entire analytic function such that $f(z_k) \neq 0$, and $a \geq 0$ and $b \geq 1$ are integers, then

$$[z^n] \frac{f(z)\{Q(z)\}^a}{(1-kz\{Q(z)\}^{k-1})^b} = \frac{f(z_k)}{2^{b/2}\,\Gamma(b/2)} \left(\frac{k}{k-1}\right)^{a+b/2} n^{(b-2)/2}\, z_k^{-n} \left(1 + \mathrm{O}\left(n^{-1/2}\right)\right).$$

[f][Svihla & Lladser (2024)].

Lastly, for our final and main result are the **well-known formulas**:

$$\mathbb{E}\big[\mathrm{IPL}(\mathbb{T}_n)\big] = \frac{[z^n]Q_u(z)}{[z^n]Q(z)};$$

$$\mathbb{V}\big[\mathrm{IPL}(\mathbb{T}_n)\big] = \frac{[z^n]Q_{uu}(z)}{[z^n]Q(z)} + \frac{[z^n]Q_u(z)}{[z^n]Q(z)} - \left(\frac{[z^n]Q_u(z)}{[z^n]Q(z)}\right)^2.$$

Sparsification of a random k-regular tree

Theorem.[g] If $\mathbb{T}_n$ is a uniformly at random k-regular tree of size n then

$$\mathbb{E}\big[\mathrm{IPL}(\mathbb{T}_n)\big] \sim \sqrt{\frac{\pi k}{2(k-1)}}\, n^{3/2} \quad \text{and}$$

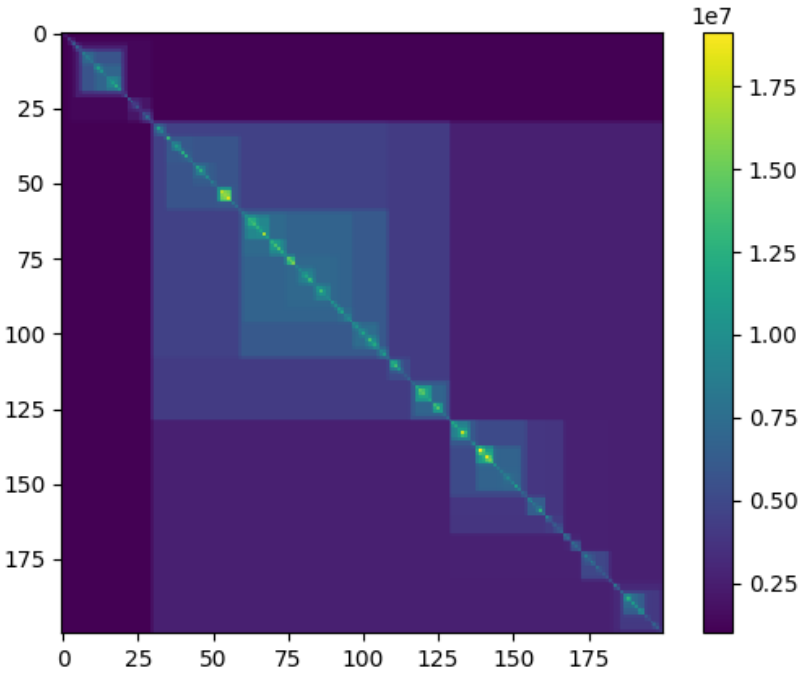
$$\mathbb{V}\big[\mathrm{IPL}(\mathbb{T}_n)\big] \sim \frac{k}{2(k-1)}\Big(\frac{10}{3} - \pi\Big) n^3.$$

In particular, if C_n and Φ_n are the phylogenetic covariance and Haar-like matrices associated with $\mathbb{T}_n$, respectively, and ζ_n denotes fraction of vanishing entries in $\Phi_n' C_n \Phi_n$ then

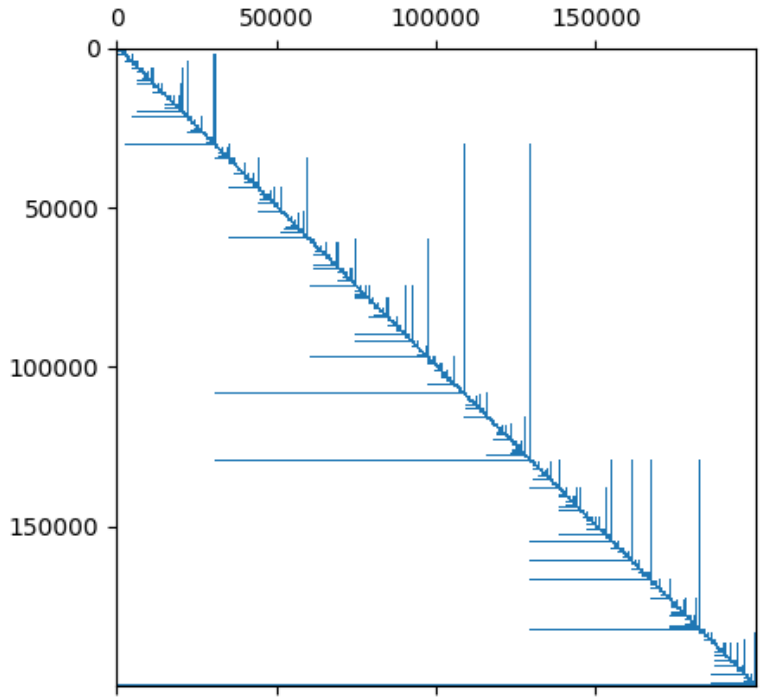
$$\lim_{n\to\infty} \zeta_n \overset{p}{=} 1.$$

[g] [Svihla & Lladser (2024)].

Does the sparsification work in practice?



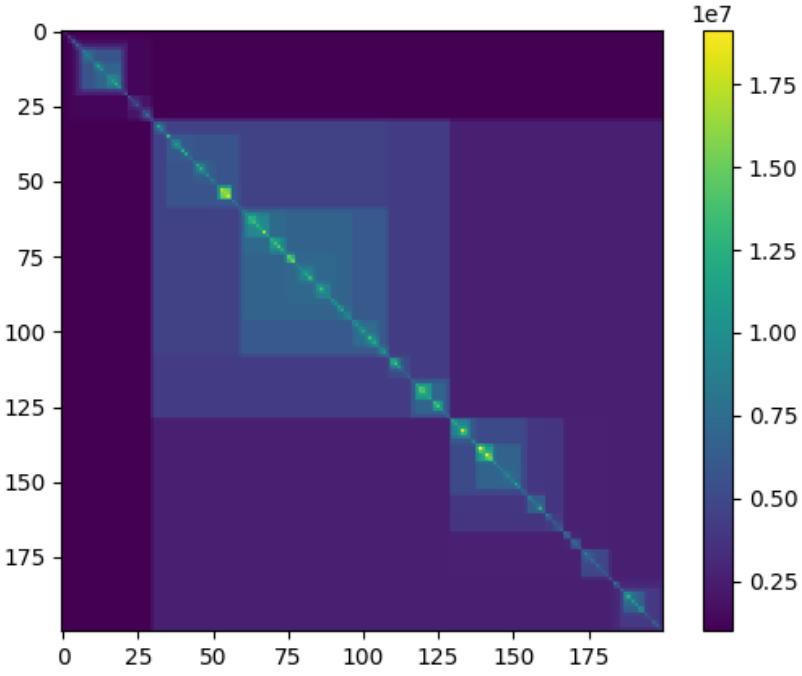
(a) 200,000×200,000 dense matrix.



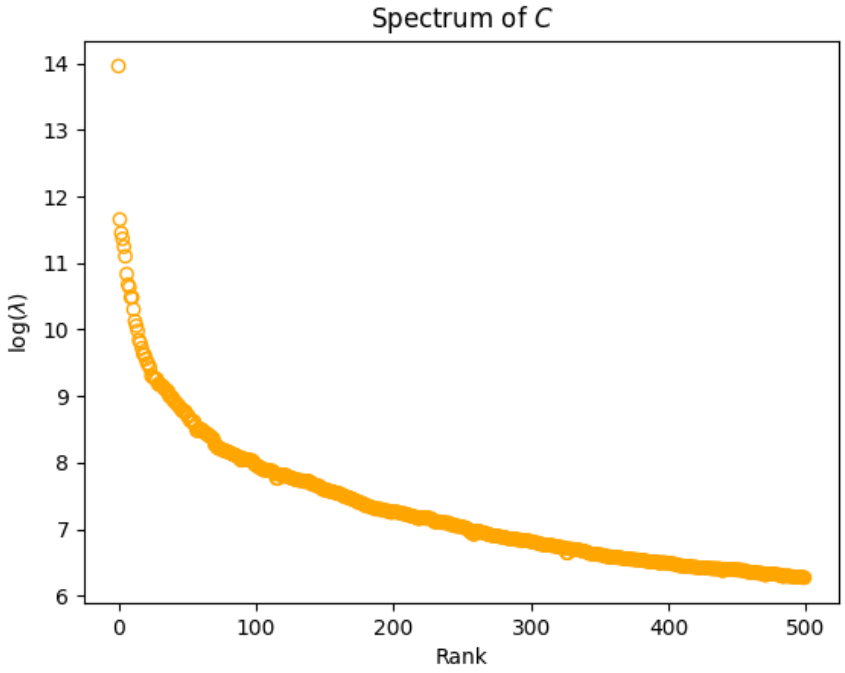
(b) Sparsified matrix.

Figure: Density/sparsity pattern of the phylogenetic covariance matrix of a random 3-regular tree with ≈200,000 leaves. The dense matrix has over 40 billion non-zero entries, but 99.97% of these vanish after changing basis to the Haar-like wavelets of the tree.

Sparsification enables manipulating large dense matrices



(a) 200,000×200,000 dense matrix.



(b) Matrix spectrum.

Figure: With over 40 billion non-zero entries, everyday software cannot manipulate the dense matrix on the left. Nevertheless, it can compute the 500 largest eigenvalues from its sparsified version, as shown in the plot on the right.

Summary

- The Haar-like wavelets can be used to sparsify phylogenetic covariance matrices.
- We can derive a lower-bound on the number of vanishing entries after changing to the Haar-like basis by counting wavelets with disjoint support.
- With high probability, a large random k-regular tree has a covariance matrix which is highly sparsified by its Haar-like basis.
- This enables manipulating large and dense phylogenetic covariance matrices from their sparsified representation.

Acknowledgements

- Manuel Lladser (Ph.D. research advisor).
- Lladser Research Group.
- Work partially funded by the NSF grant No. 1836914.

Thank you!