Atypical decay rates for atypical heights in RRT

Heng Ma 馬恒 (Technion)

Joint work (in progress) with

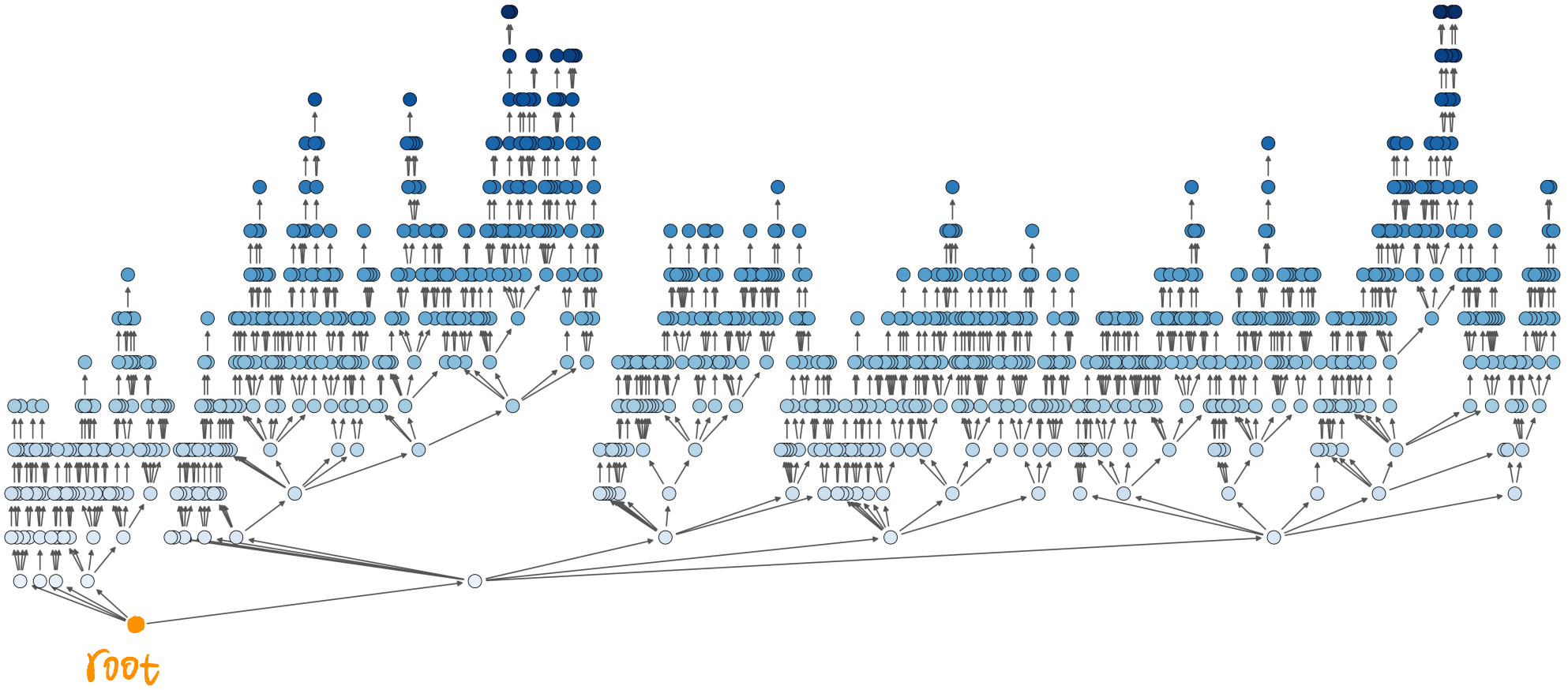
Xinxin Chen 陈昕昕

(Beijing Normal University)

Random recursive tree (RRT) T_n

- Start with a root v_1
- At each step $n \geq 2$, introduce a new vertex v_n.
- We attach v_n to an existing vertex, choosen uniformly at random from $\{v_1 \cdots v_{n-1}\}$.

Also known as the "Uniform attachment tree".



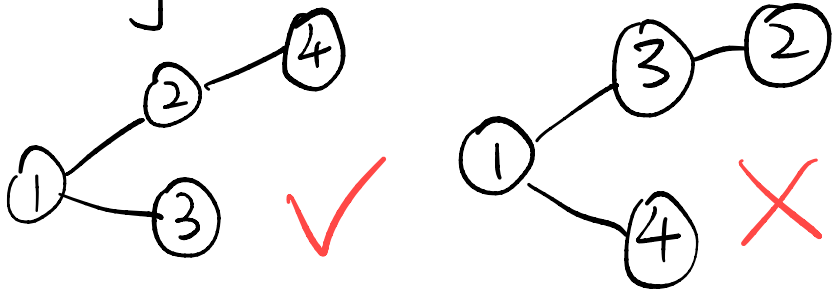
RRT on 1000 nodes

As uniform distribution on increasing trees

- A tree on $\{v_1, \cdots, v_n\}$ is called an increasing tree if

  1) The root has smallest label v_1

  2) The label along any path starting from the root is increasing



- For any increasing tree T,

$$\mathbb{P}(\mathcal{T}_n = T) = \frac{1}{(n-1)!}$$

Remark (General attachment rules)

- Start with v_1.
- At each step $n \geq 2$, introduce a new vertex v_n.
- Attach v_n to an existing vertex $v_{I_n} \in \{v_1 \cdots v_{n-1}\}$, where

$$\mathbb{P}(I_n = j \mid \mathcal{T}_{n-1}) \propto \theta \cdot \deg^{+}_{\mathcal{T}_{n-1}}(v_j) + 1,$$

$\deg^{+}_{\mathcal{T}_{n-1}}(v_j)$: out degree of v_j

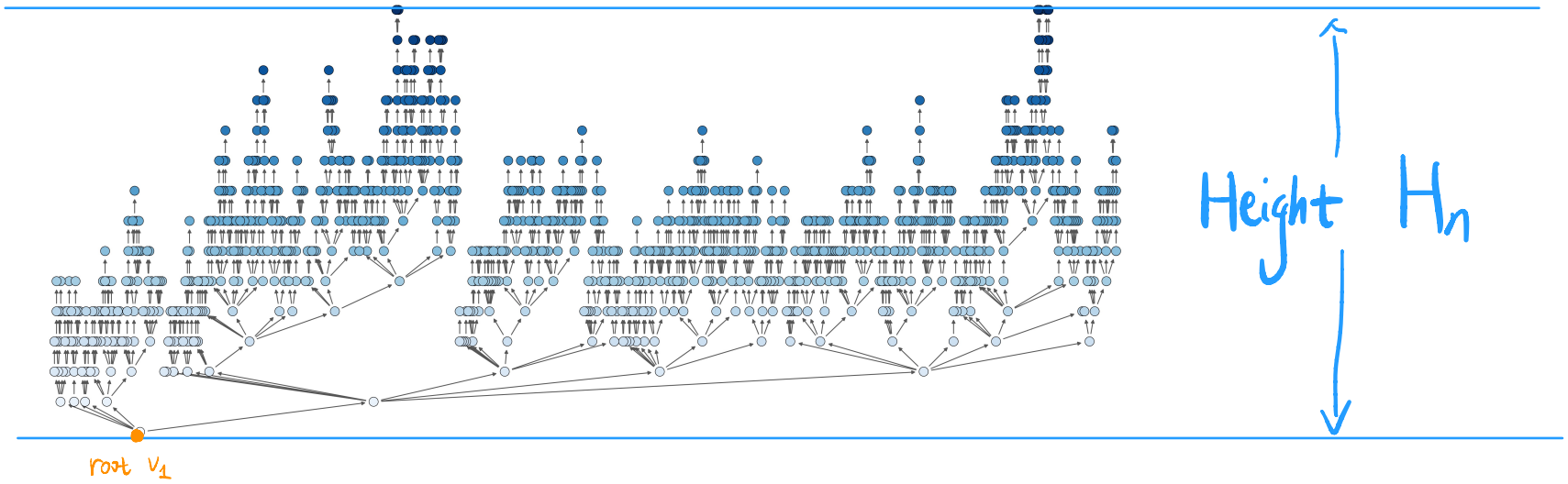
$\theta = 0$	RRT
$\theta = 1$	preferential attachment tree / Plane Oriented Recursive Tree
$\theta = -\frac{1}{2}$	Binary Search tree

Typical Behaviors of RRT

The Height of RRT

- Denote by H_n the height of the tree $\mathcal{T}_n$:

$$H_n := \max_{1 \le j \le n} d(v_1, v_j)$$



- Szymanski '90 : with high prob. $1 \le \frac{H_n}{\log n} \le e$.
- Pittel '94 proved that $\lim_{n\to+\infty} \frac{H_n}{\log n} = e$ almost surely, by using a Continuous-time Embedding method
- Addario-Berry - Ford '13 :

$$\begin{cases} \mathbb{E} H_n = e \log n - \frac{3}{2} \log\log n + O(1) \\ (H_n - \mathbb{E} H_n)_{n\ge 1} \text{ is tight} : \ \forall\, \lambda \in (0, \frac{1}{2e}), \ n \ge 1 \end{cases}$$

$$\mathbb{P}\left(|H_n - \mathbb{E} H_n| > k \right) \lesssim_\lambda e^{-\lambda k}$$

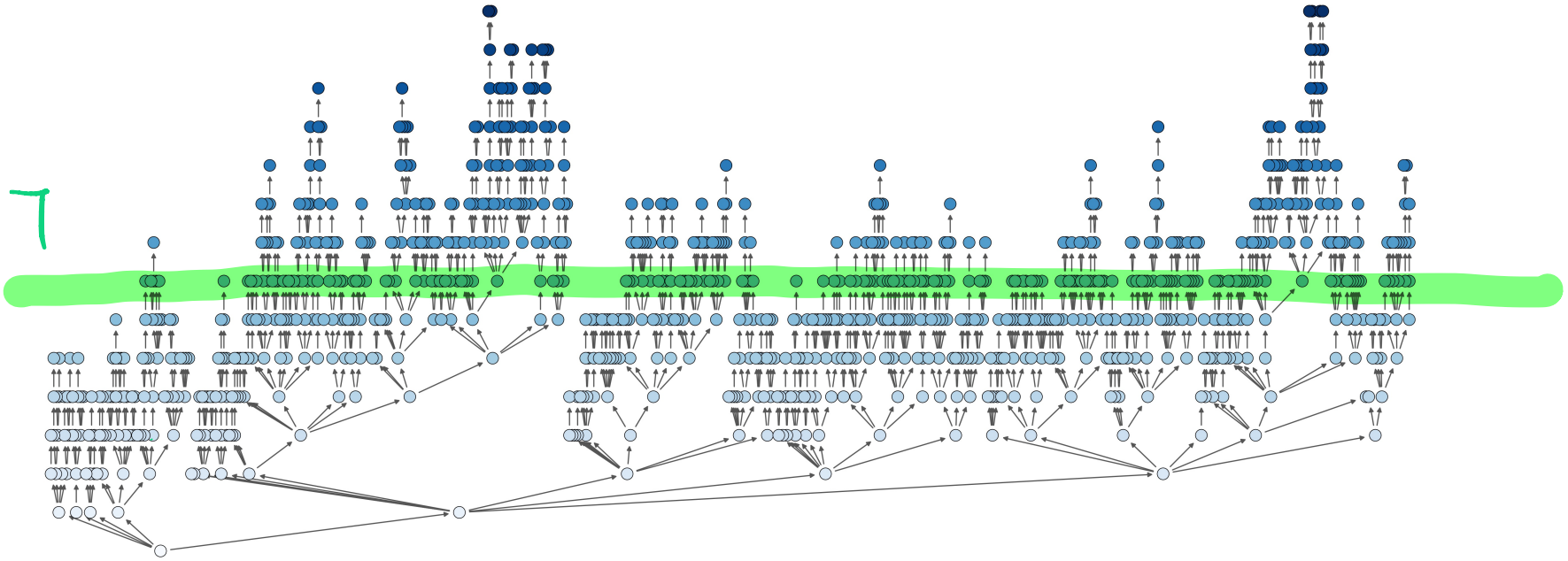
Level Sets of RRT

- Let $X_{n,k} :=$ number of nodes at level k in $\mathcal{T}_n$

$$= \sum_{j=1}^{n} \mathbb{1}\{ d(v_1, v_j) = k \}$$

$(X_{n,k})_{k \geq 1}$ is also called the profile of $\mathcal{T}_n$ in the literature.

Level ℓ



Fuchs-Hwang-Neininger '06:

- For $0 \le \alpha < e$, if $k = k(n)$ satisfies $\frac{k}{\log n} \to \alpha$ then

$$\mathbb{E}X_{n,k} = n^{\alpha(1-\log\alpha)+o(1)}, \quad \frac{X_{n,k}}{\mathbb{E}X_{n,k}} \xrightarrow{d} X_\alpha$$

where X_α satisfies

$$X \overset{d}{=} \alpha U^\alpha X + (1-U)^\alpha \tilde{X}$$

$\tilde{X}$: an i.i.d. copy of X

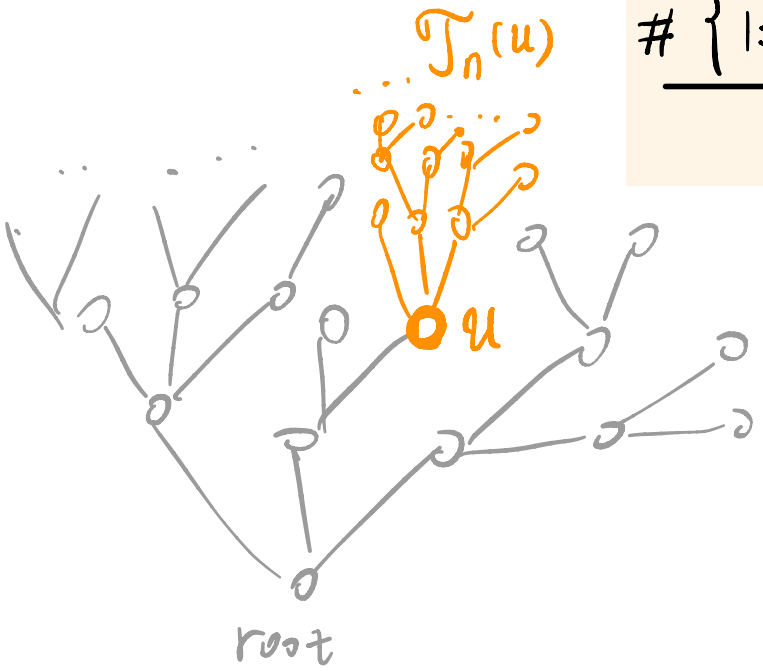
$U \sim \text{Unif}[0,1]$

- CLT when $\alpha = 0$: If $k = k(n)$ satisfies $\frac{k}{\log n} \to 0$ then

$$\frac{(k-1)!\sqrt{2k-1}}{(\log n)^{k-\frac{1}{2}}}\left[X_{n,k} - \frac{(\log n)^k}{k!}\right] \xrightarrow[n\to+\infty]{d} \text{Normal}(0,1)$$

Limiting Fringe Tree of RRT

Aldous '91: Uniformly choose a node u in $\mathcal{T}_n$. What does the subtree of u looks like?



$$\frac{\#\left\{1 \le j \le n : \mathcal{T}_n(u) \unlhd \tau\right\}}{n} \xrightarrow{\text{a.s.}} \mathbb{P}(\mathcal{T}_M \unlhd \tau)$$

where M is a R.V. independent of $(\mathcal{T}_n)$, with $\mathbb{P}(M \ge n) = \frac{1}{n}$.

Degree Profile of RRT

As a consequence,

$$\frac{\#\left\{1 \le j \le n : \deg_{\mathcal{T}_n}(v_j) = k\right\}}{n} \xrightarrow[n \to +\infty]{a.s.} 2^{-k}$$

In particular,

proportion of leaves $\longrightarrow$ 50%

Our focus:

Rare events concerning RRT height

RRT being atypically short

Fix $\alpha \in (0,1)$, what is the decay rate of

$$\mathbb{P}(H_n \leq \alpha e \log n)$$

RRT being atypically tall

Fix $\beta > 1$, what is the decay rate of

$$\mathbb{P}(H_n \geq \beta e \log n)$$

Theorem 1 (Chen-M.' 26+)

Define $m_\alpha(n) := n^{1-\alpha}(\log n)^{-\frac{3}{2e}}$ and $\omega_\alpha(n)$ as follows

$$\mathbb{P}(H_n \geq \alpha e \ln n) = e^{-\omega_\alpha(n)\, m_\alpha(n)}$$

Then for $\alpha < 1$,

$$\liminf_{n\to+\infty} \omega_\alpha(n) = +\infty$$

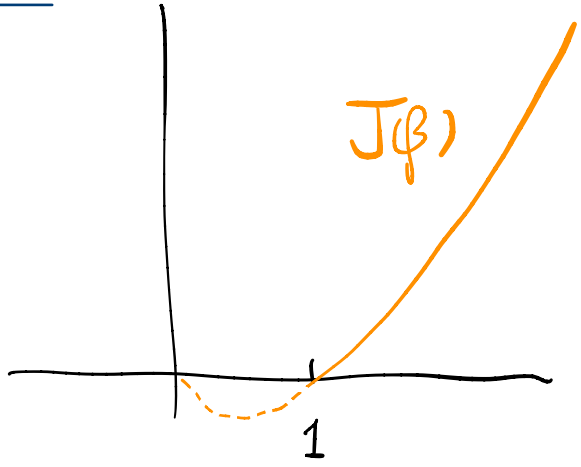
Moreover, for any integer $k \in \mathbb{N}$,

$$\limsup_{n\to+\infty} \frac{\omega_\alpha(n)}{\log^{(k)}(n)} < +\infty .$$

$\log^{(k)}(x) := \log^{(k-1)}(\log x)$

Theorem 2 (Chen-M.'26+)

- Define $J(\beta) := e \cdot \beta \ln \beta$



Then for $\beta > 1$, we have

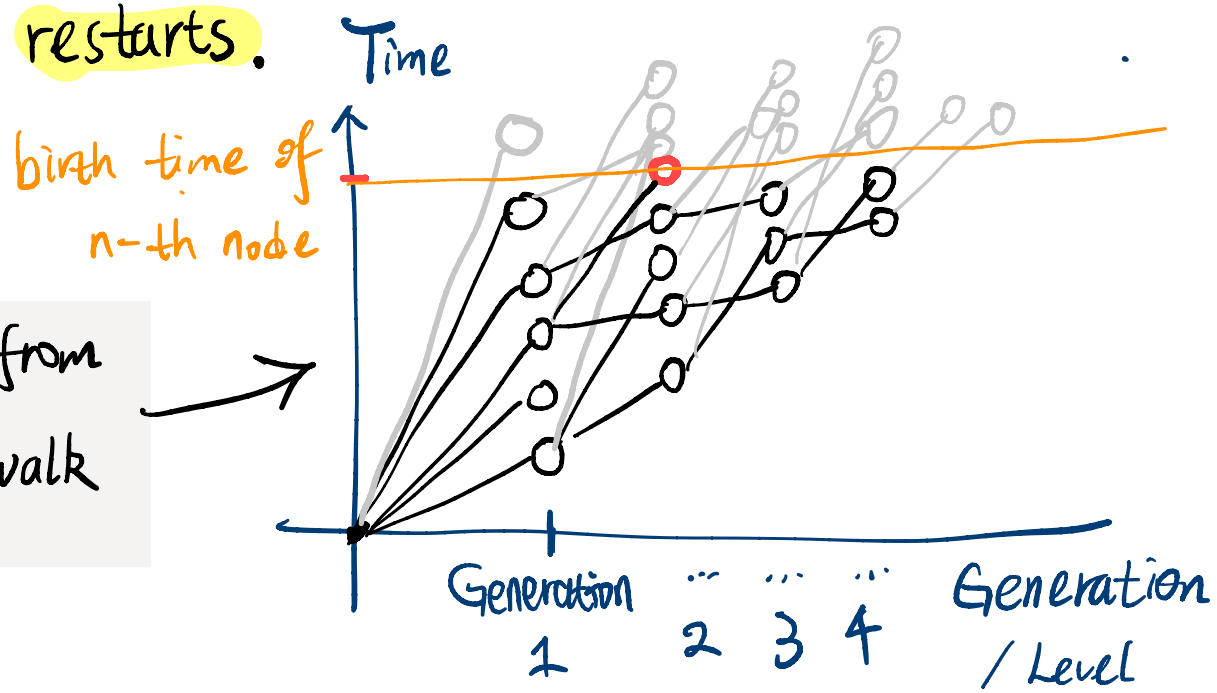
$$\mathbb{P}(H_n > \beta e \log(n)) = n^{-J(\beta)+o(1)} = e^{-[J(\beta)+o(1)]\log n}$$

A powerful approach

Pittel's Continuous-time embedding

- Each node has an independent alarm clock, exponentially distributed with rate 1
- When it rings, this node give birth to a new child and its clock restarts.

Extracting a RRT from a branching random walk →



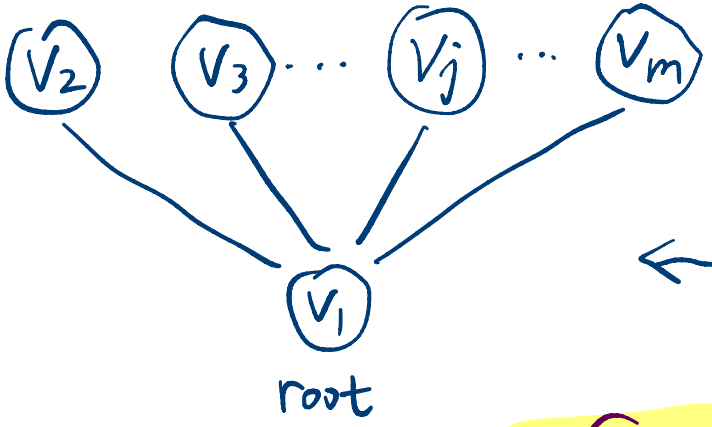
Proof ideas of Thm 1

Strategy to make

the tree atypically short

A crude strategy:

Make the tree as short as possible at an early stage, then let it grow "freely".



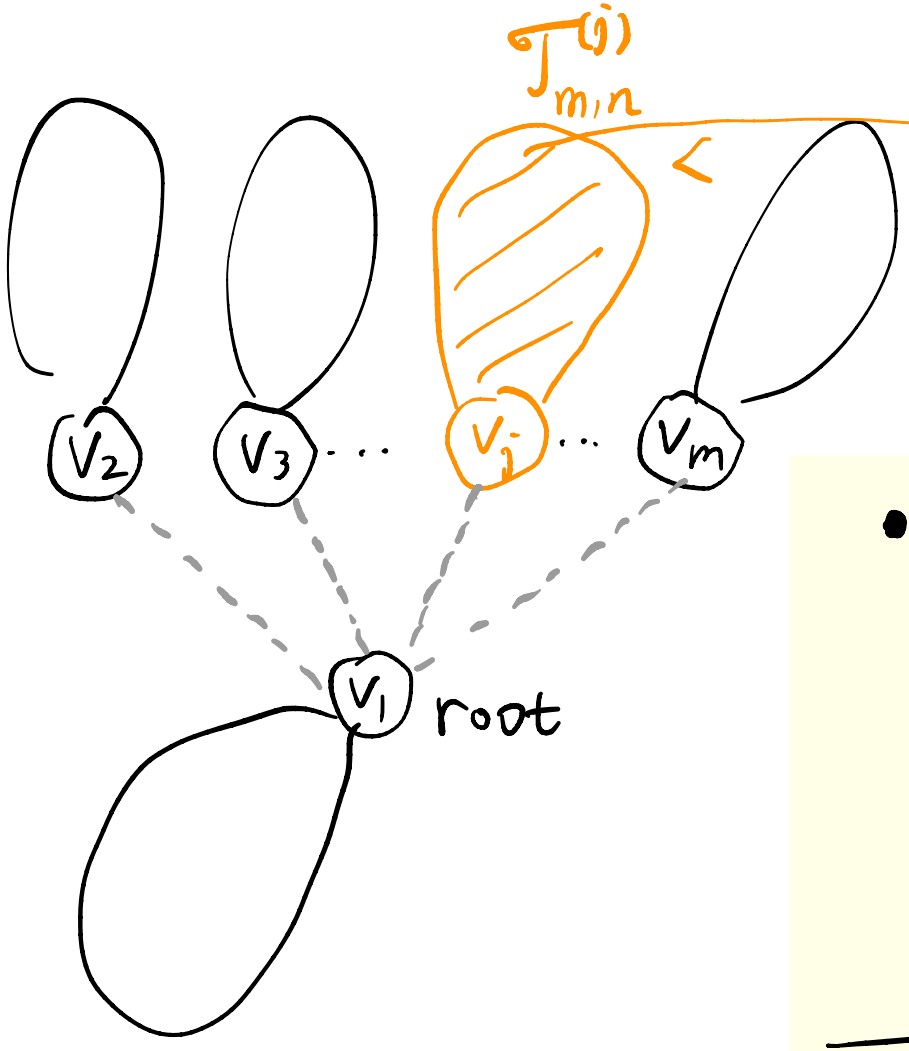
That is, fix $m \geq 2$,

we ask $H(T_m) = 1$

Cost: $\mathbb{P}(H(T_m) = 1)$

$$= \frac{1}{(m-1)!} = e^{-m \ln m + O(m)}$$

Make the tree as short as possible at an early stage, then let it grow "freely".



- For each $2 \le j \le m$, let $T^{(j)}_{m,n}$ denote the tree containing V_j after removing edges in T_m.

- Then

$$\begin{cases} H(T_m) = 1 \\ \max\limits_{2 \le j \le m} H(T^{(j)}_{m,n}) \le 2e\log n - 1 \end{cases}$$

$$\Rightarrow \quad H_n := H(T_n) \le 2e \log n$$

Observation 1: Conditionally on $\left\{\mathcal{T}_m = \tau,\ |\mathcal{T}_{m,n}^{(j)}| = n_j,\ 1 \le j \le m\right\}$

$\mathcal{T}_{m,n}^{(j)}$ are independent RRT on n_j nodes.

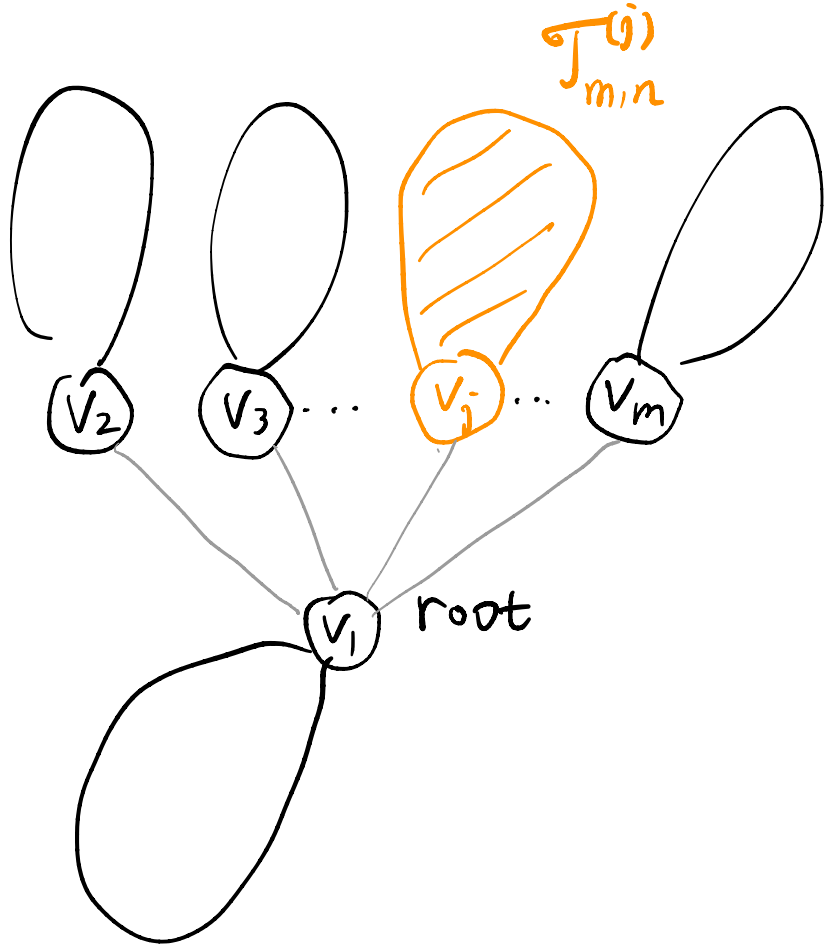
Observation 2:

$\left(|\mathcal{T}_{m,n}^{(j)}| : 1 \le j \le m\right)$

has the Uniform distribution on

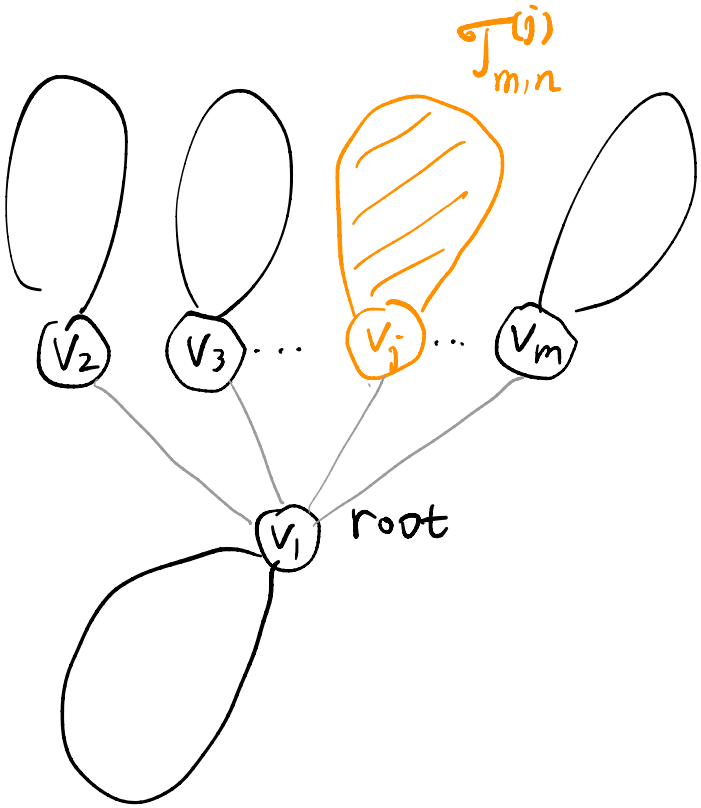
$$\left\{(n_j)_{j=1}^m : n_j \in \mathbb{N};\ \sum_{j=1}^m n_j = n\right\}.$$

In particular, $\mathbb{E}|\mathcal{T}_{m,n}^{(j)}| = \frac{n}{m}$.



- Need a lower bound for

$$\mathbb{P}\left(\max_{2\le j\le m} H(T^{(j)}_{m,n}) \le 2e\log n - 1\right)$$



- Suppose $|T^{(j)}_{m,n}| \le 100\frac{n}{m}$, then

$$H(T^{(j)}_{m,n}) \overset{st}{\le} H_{100\frac{n}{m}} = \mathbb{E}[H_{\frac{n}{m}}] + O_{\mathbb{P}}(1)$$

- Choose m carefully so that $m = C_{st} \times m_2(n)$

$$\mathbb{E}[H_{\frac{n}{m}}] = 2e\log n - \text{large constant}$$

and hence

$$\mathbb{P}\left(H_{100\frac{n}{m}} \le 2e\log n - 1\right) \ge \frac{1}{2}.$$

All together.

$$\mathbb{P}(H_n \le 2e \ln n) \ge \mathbb{P}(H(T_m) = 1) \frac{1}{(m-1)!} = e^{-m \ln m + O(m)}$$

$$\times \mathbb{P}\left(|T^{(i,j)}_{m,n}| \le 100 \frac{n}{m} \quad \forall\, 1 \le i \le j \le m\right)$$

One can show it's greater than e^{-cm}

$$\times \mathbb{P}(H_{100\frac{n}{m}} \le 2e \log n - 1)^m$$

By our choice of m, it is greater than 2^{-m}

$$\ge e^{-c\, m \ln m} \ge e^{-c\, m_2(n) \log n}$$

□

All together,

$$\mathbb{P}(H_n \le 2e \ln n) \ge \mathbb{P}(H(T_m) = 1) \cdot \frac{1}{(m-1)!} = e^{-m \ln m + O(m)}$$

$$\times \mathbb{P}\left(|T^{(i,j)}_{m,n}| \le 100\frac{n}{m} \quad \forall\, 1 \le i \le j \le m\right)$$

One can show it's greater than e^{-m}

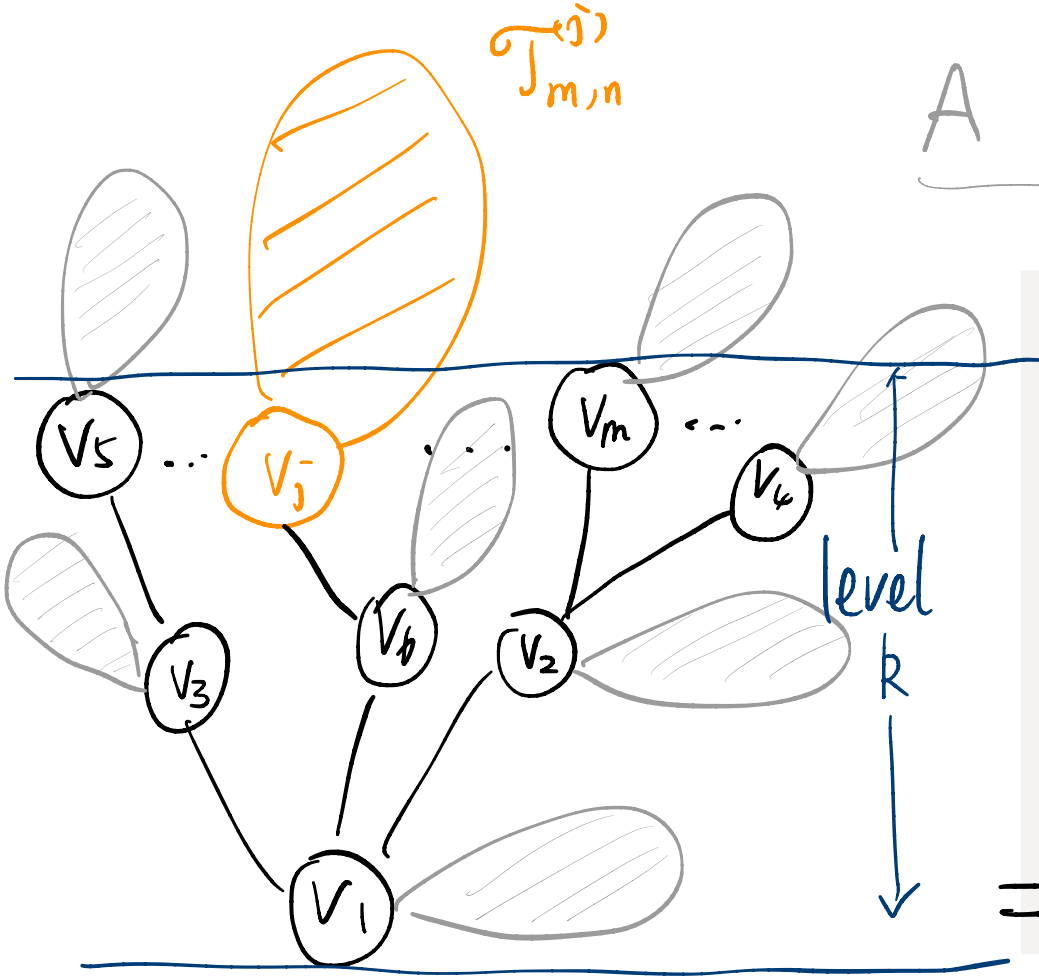
$$\times \mathbb{P}\left(H_{100\frac{n}{m}} \le 2e \log n - 1\right)$$

by our choice of m, it is greater than $\frac{1}{2}$

$$\ge e^{-c\, m \ln m} \ge e^{-c\, m_2(n) \log n}$$

Just need to improve the crude requirement $H(T_m)=1$

□



A slight improvement:

Fix any $k \in \mathbb{N}$. Then

$$\begin{cases} H(T_m) = k \\ \max\limits_{1 \le j \le m} H(T^{(j)}_{m,n}) \le 2e\log n - k \end{cases}$$

$$\Rightarrow \quad H_n := H(T_n) \le 2e \log n$$

- $\mathbb{P}(H(T_m) = k) = \dfrac{A_k(m)}{(m-1)!}$ ← # Increasing trees with height k on m nodes.

$$\ge e^{-m \log^{(k)} m} = e^{-m \log^{(k)} n}$$

□

Remarks

- We have a intuitive enumeration showing that

$$A_k(n) \geq \exp(n\log n - n\log^{(k)} n + O(n)) \qquad (*)$$

- It is well known that, the generating function

$$F_k(z) = \sum_{n=1}^{\infty} \frac{A_k(n)}{n!} z^n = \sum_{n=1}^{\infty} \frac{1}{n} \mathbb{P}(H_n \leq k)\, z^n$$

satisfies

$$\begin{cases} F_k'(z) = \exp(F_{k-1}(z)) \\ F_1(z) = e^z - 1 \,;\quad F_k(0) = 0 \end{cases}$$

In principle, $(*)$ should be proved by analyzing $F_k(z)$

But it seems hard (for us) to do this ...

Proof ideas of Thm 1

Constraints on

many sub-structures

$m = \frac{m_2(n)}{\text{large Cst}}$

(1) Reselect m so that

$$\mathbb{E}[H_{\frac{1}{4}\frac{n}{m}}] = 2e\log n + \text{large constant}$$

and hence

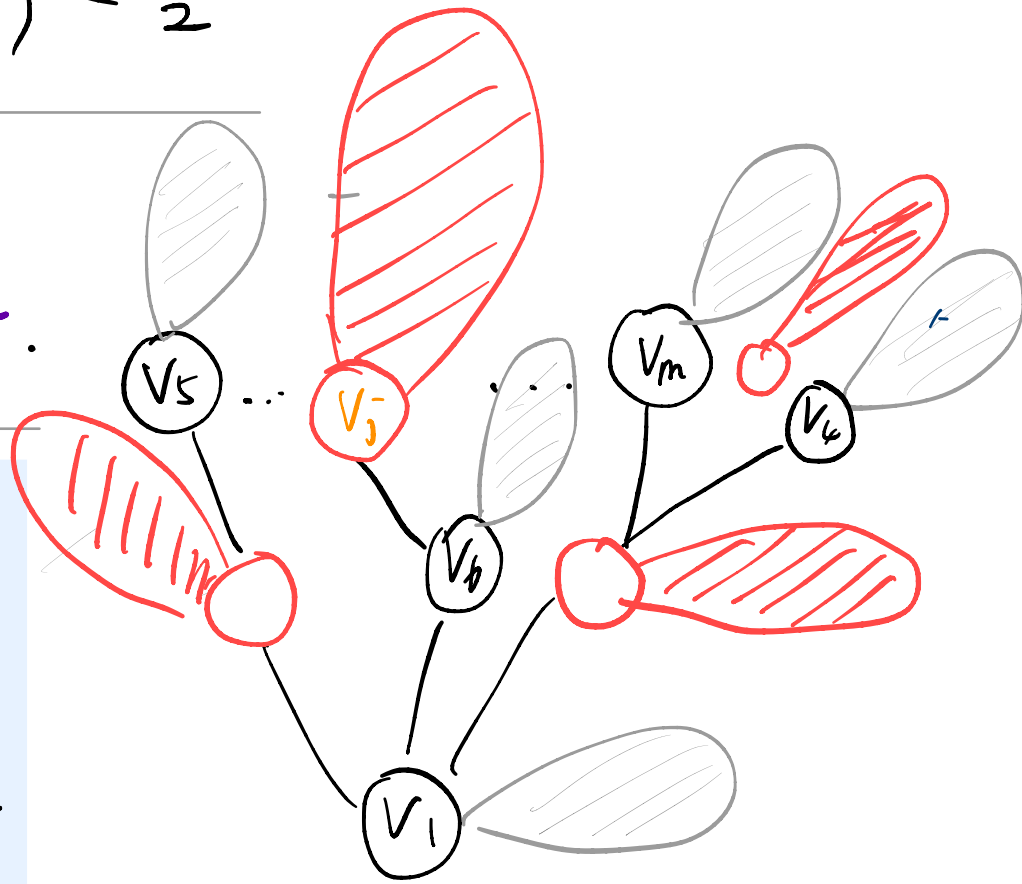
$$\mathbb{P}\left(H_{\frac{1}{4}\frac{n}{m}} \le 2e\log n\right) < \frac{1}{2}$$

---

(2) Let $BIG_{m,n}$ be the collection of all $T^{(j)}_{m,n}$'s with $|T^{(j)}_{m,n}| \ge \frac{1}{4}\frac{n}{m}$.

Combing (2) with (1):

$$\mathbb{P}(H_n \le 2e\log n) \le \mathbb{E}\left[2^{-|BIG_{m,n}|}\right].$$



- $\mathbb{P}(H_n \le 2e\log n) \le \mathbb{E}\left[2^{-|\mathrm{BIG}_{m,n}|}\right]$

CLAIM: $\mathbb{P}(|\mathrm{BIG}|_{m,n} < \gamma m) \le e^{-C^*_\gamma m}$

$\le 2^{-cm} + e^{-c'm} \le e^{-c''m}$ $\square$

Recall that $\left(|\mathcal{T}^{(j)}_{m,n}|\right)_{j=1}^{m}$ $\overset{\text{Unif}}{\sim}$ $\left\{(n_j)_1^m : n_j \in \mathbb{N}, \sum_1^m n_j = n\right\}$

$\xi_j := \mathbb{1}\{\mathcal{T}^{(j)}_{m,n}$ is small$\}$ are negative correlated

$\mathbb{P}\left(\sum_1^m \xi_j \ge (1-\gamma)m\right) \le \mathbb{P}\left(\sum_1^m \xi_j \overset{\text{IID}}{\ge} (1-\gamma)m\right)$

$\lesssim \exp\{-D(1-\gamma \| p_0)\, m\}$

[Rmk: $\lim_{\gamma \downarrow 0} C^*_\gamma = +\infty$]

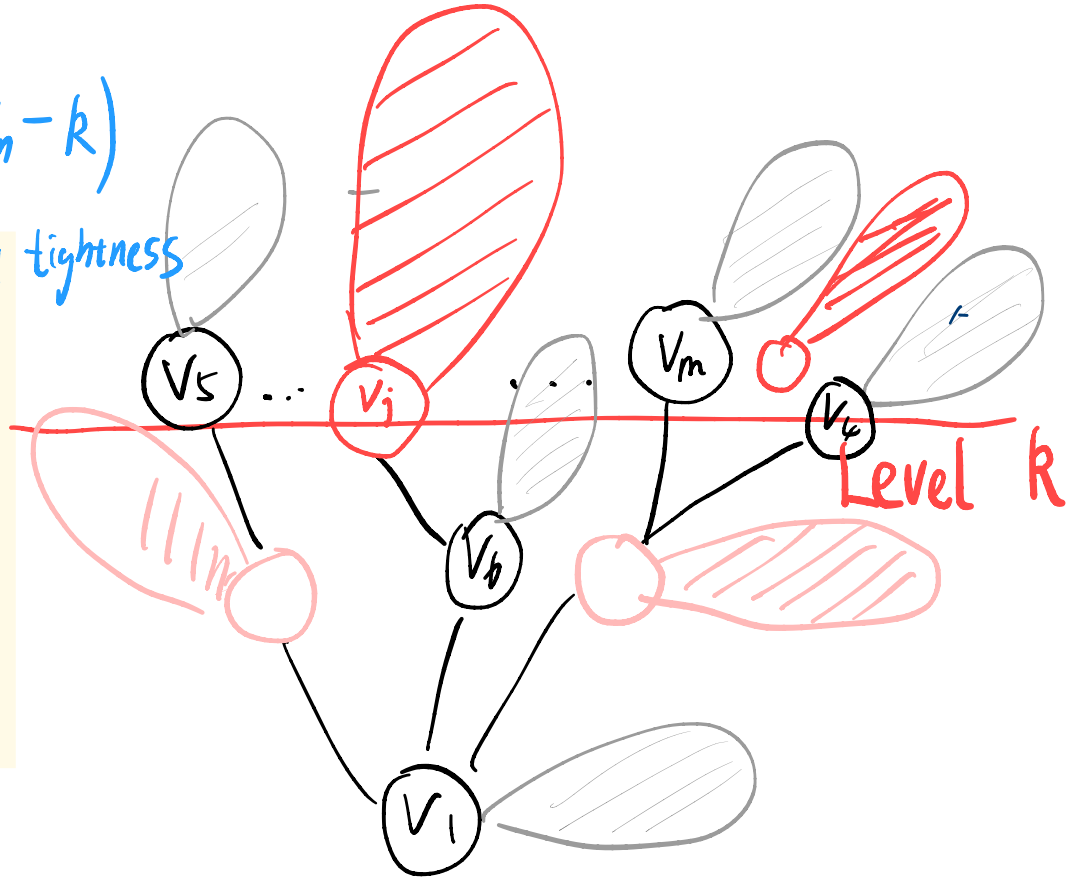
Where can be improved ?

- Fix $k \in \mathbb{N}_+$
- Let $BIG^{(k)}_{m,n} = \left\{ 1 \le j \le m : \boxed{d(v_1, v_j) \ge k} \right.$
  $\left. \overrightarrow{T}^{(j)}_{m,n} \text{ is big; i.e., } |\overrightarrow{T}^{(j)}_{m,n}| \ge \frac{1}{4}\frac{n}{m} \right\}$

Then

$$\delta(k) = \sup_{n \ge 1} \mathbb{P}\left(H_n \le \mathbb{E}H_n - k\right) \xrightarrow{k \to +\infty} 0 \text{ by tightness}$$

$$\mathbb{P}(H_n \le 2e \ln n) \le \mathbb{E}\left[\delta(k)^{|BIG^{(k)}_{m,n}|}\right]$$



✡ Then for any $\varepsilon \in (0,1]$, $\exists\, c_\varepsilon > 0$ s.t. for $n \geq 1$

$$\mathbb{P}\left(\sum_{l=0}^{k} X_{n,l} > \varepsilon n\right) \leq e^{-c_\varepsilon n \log^{(k)}(n)}$$

↖ size of level l in $\mathcal{T}_n$

- Consequently, $\mathbb{P}\left(|BIG_{m,n}^{(k)}| < \frac{\gamma}{2} m\right)$

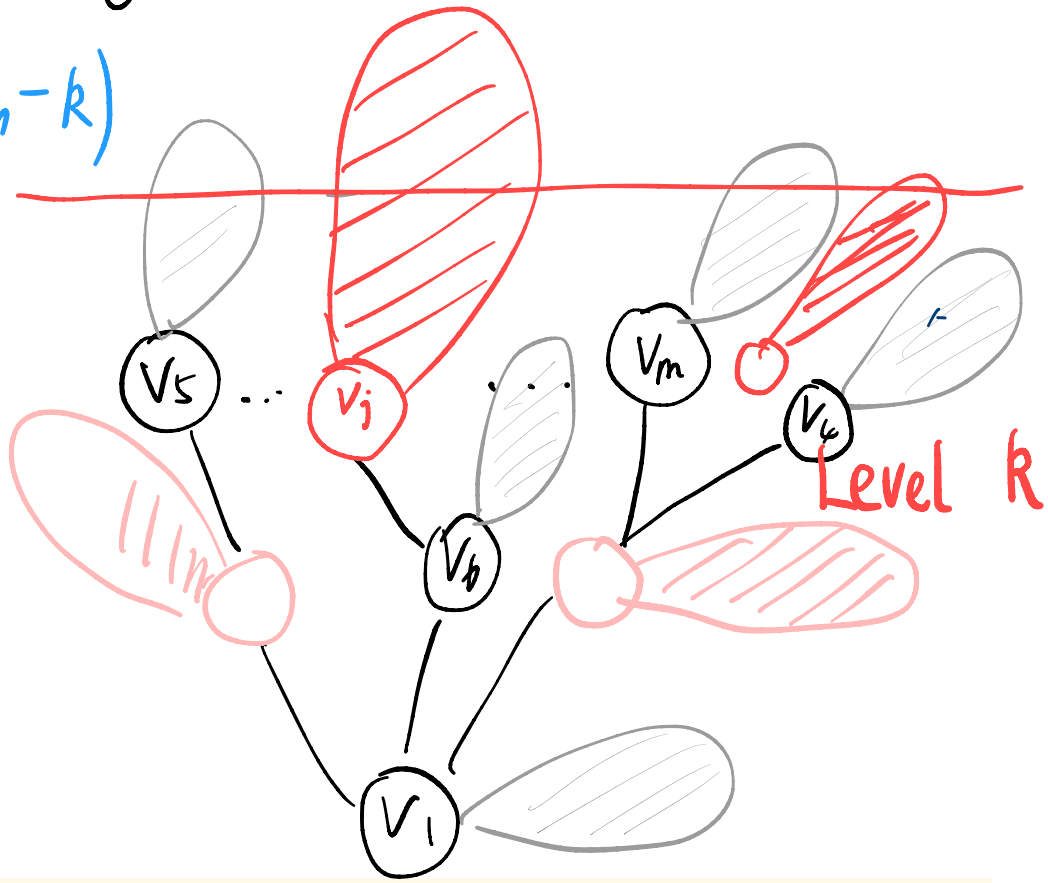
$$\leq \mathbb{P}\left(|BIG|_{m,n} < \gamma m\right) + \mathbb{P}\left(\sum_{l=0}^{k} X_{m,l} > \frac{\gamma}{2}\right)$$

$$\lesssim e^{-c_\gamma^* m} + e^{-\Theta(m \log^{(k)} m)}$$

$$BIG^{(k)}_{m,n} = \left\{ 1 \le j \le m : \; d(v_1, v_j) \ge k, \; \mathcal{T}^{(j)}_{m,n} \text{ is big}, \; |\mathcal{T}^{(j)}_{m,n}| \ge \frac{1}{4}\frac{n}{m} \right\}$$

Then

$$\delta(k) = \sup_{n \ge 1} \mathbb{P}\left(H_n \le \mathbb{E}H_n - k \right)$$



$$\mathbb{P}(H_n \le 2e \ln n)$$

$$\le \mathbb{E}\left[\delta(k)^{|BIG^{(k)}_{m,n}|} \right]$$

$$\le \delta(k)^{\gamma m} + e^{-C^*_\gamma m}$$

$$\le e^{-Km}$$ for any $K > 0$, provided γ small then k large enough. $\square$

Future directions

- Identifying the decay rate of $w_2(n)$
- LDP on the height of Preferential attachment tree

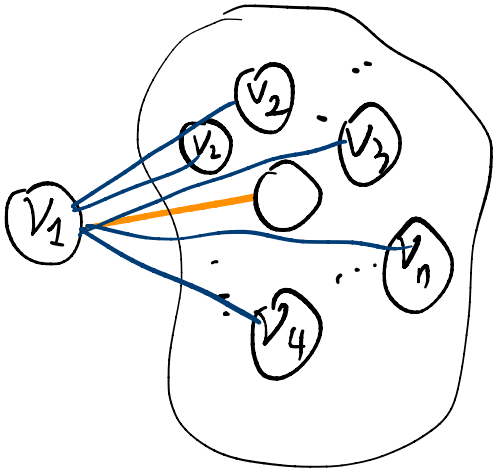
  binary search tree and other random tree models.
-

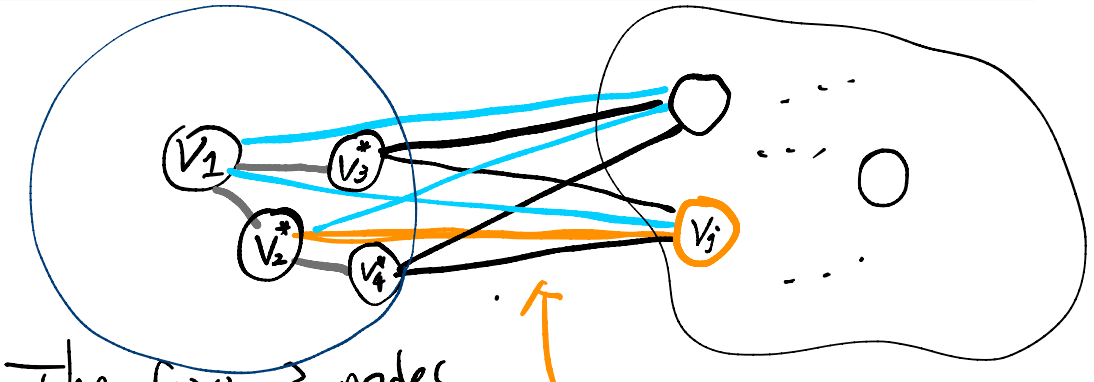

Thanks for
your attention !

As a shorest-path tree

- K_n = Complete graph on $\{v_1 \cdots v_n\}$
- assign IID exponential weights (mean 1) to each edge

The tree $\mathbb{T}_n$, formed by the shorest-path from v_1 to all other nodes $\{v_2 \cdots v_n\}$, is a RRT.





The frist 3 nodes closest to v_1

Find the 4th closest node closest to v_1