Large fringe and non-fringe subtrees in conditional Galton-Watson trees

Xing Shi Cai, Luc Devroye

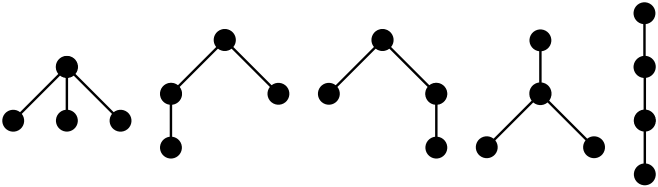
School of Computer Science
McGill University

Probabilistic Midwinter Meeting
Umeå University
Jan 18, 2017

Outline

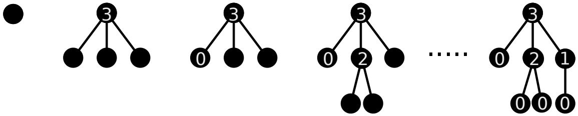
What is a tree

- A tree is an *acyclic graph*.
- In this talk, trees are *unlabeled*, *rooted*, and *ordered* (plane trees).



Galton-Watson trees

- A Galton-Watson (GW) tree $\mathcal{T}^{gw}$ starts with a single node.
- Each node in $\mathcal{T}^{gw}$ chooses a random number of child nodes independently from the same distribution ξ.
- Introduced by Bienaymé, 1845.



Note

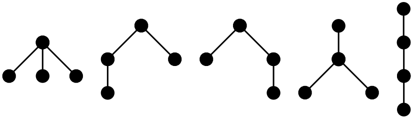
We will always assume that $\mathbb{E}\xi = 1$ and $\operatorname{Var}(\xi) \in (0, \infty)$.

Conditional Galton-Watson trees

- A conditional GW Tree $\mathcal{T}_n^{gw}$ is $\mathcal{T}^{gw}$ restricted to $|\mathcal{T}^{gw}| = n$.
- So $\mathbb{P}\left\{\mathcal{T}_n^{gw} = T\right\} = \mathbb{P}\{\mathcal{T}^{gw} = T \mid |\mathcal{T}^{gw}| = n\}$.
- It covers many uniform random tree models:
  - full binary trees
  - binary trees
  - d-ary trees
  - Motzkin trees
  - Plane trees
  - Cayley trees

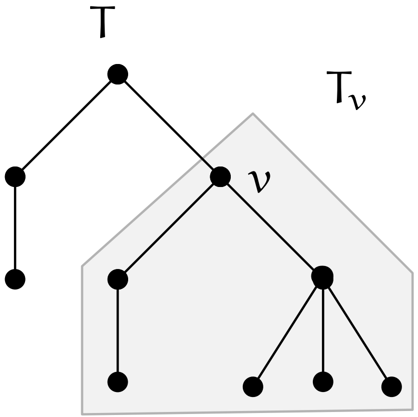
Example of conditional Galton-Watson trees

- Let $\mathbb{P}\{\xi = i\} = 1/2^{i+1}$.
- In other words, $\xi \overset{\mathcal{L}}{=} \mathrm{Ge}(1/2)$.
- $\mathcal{T}_n^{gw}$ is uniformly distributed among all trees of size n.

$\mathbb{P}\{\mathcal{T}^{gw} = \mathsf{T}\} = 2^{-7}$ for $\mathsf{T} \in \Bigg\{$  $\Bigg\}$

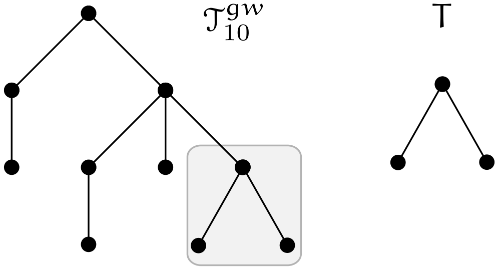
Fringe subtrees

- For a node v of a tree T, the *fringe subtree* T_v contains v and all its decedents.
- It is what normally called a "subtree".



Fringe subtree count

- Let $N_T(\mathcal{T}_n^{gw})$ be the number of fringe subtrees of shape T in $\mathcal{T}_n^{gw}$.



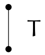
$$N_T(\mathcal{T}_{10}^{gw}) = 1$$

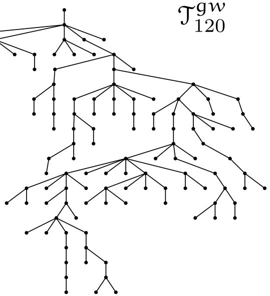
Fringe subtree count: bigger example

- In the next example,

$$\frac{N_{\mathsf{T}}(\mathcal{T}_n^{gw})}{n} = \frac{15}{120} = \frac{1}{8} = \pi(\mathsf{T}) \equiv \mathbb{P}\{\mathcal{T}^{gw} = \mathsf{T}\}.$$

- Is this just a coincidence?





$\xi \sim \mathrm{Ge}(1/2)$

$N_{\mathsf{T}}(\mathcal{T}^{gw}_{10}) = 15$

$\pi(\mathsf{T}) = 1/8$

What is known

- For large n, fringe subtrees in $\mathcal{T}_n^{gw}$ behave like independent copies of $\mathcal{T}^{gw}$.
- Take a uniform random fringe subtree of $\mathcal{T}_n^{gw}$, the probability to get T is about $\pi(T) \equiv \mathbb{P}\{\mathcal{T}^{gw} = T\}$.
- So $N_T(\mathcal{T}_n^{gw}) \approx \mathrm{Bi}(n, \pi(T))$.

What is known cont.

Theorem Aldous (1991) (Law of large number)

As $n \to \infty$,

$$\frac{N_T(\mathcal{T}_n^{gw})}{n} \xrightarrow{p} \pi(T).$$

Theorem Janson (2016) (Central limit theorm)

As $n \to \infty$,

$$\frac{N_T(\mathcal{T}_n^{gw}) - n\pi(T)}{\gamma\sqrt{n}} \xrightarrow{d} N(0,1),$$

where γ *is a constant.*

What do we want to know

- What if the T in $N_T(\mathcal{T}_n^{gw})$ changes with n?
- The height of the largest complete r-ary fringe subtree.
- The largest k such that $\mathcal{T}_n^{gw}$ contains all trees of size $\leqslant k$ as fringe subtree.
- What about non-fringe subtrees?

Outline

Large fringe subtrees

- If $|T_n| \to \infty$, then $\pi(T_n) \equiv \mathbb{P}\{\mathcal{T}^{gw} = T_n\} \to 0.$
- Then we should have

$$N_{T_n}(\mathcal{T}_n^{gw}) \approx \mathrm{Bi}(n, \pi(T_n)) \approx \mathrm{Po}(n\pi(T_n)).$$

Theorem 1.2

Let $k_n = o(n)$ *and* $k_n \to \infty$. *Then*

$$\lim_{n\to\infty} \sup_{T:|T|=k_n} d_{TV}\left(N_T(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(T))\right) = 0.$$

Large fringe subtrees cont.

Theorem 1.2 cont.

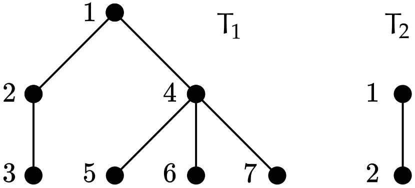
So letting $(T_n)_{n\geqslant 1}$ *be a sequence of trees with* $|T_n| = k_n$*, we have:*

1. *If* $n\pi(T_n) \to 0$*, then* $N_{T_n}(\mathcal{T}_n^{gw}) = 0$ *whp.*
2. *If* $n\pi(T_n) \to \mu \in (0, \infty)$*, then* $N_{T_n}(\mathcal{T}_n^{gw}) \xrightarrow{d} \mathrm{Po}(\mu)$.
3. *If* $n\pi(T_n) \to \infty$*, then*

$$\frac{N_{T_n}(\mathcal{T}_n^{gw}) - n\pi(T_n)}{\sqrt{n\pi(T_n)}} \xrightarrow{d} N(0,1).$$

The degree sequence

- The *degree* of a node is the number of its children.
- The *degree sequence* of a tree, is the list of degrees of its nodes in Depth-First-Search order.
- We can count fringe subtree through degree sequence.



Degree sequence: $(2, \underline{1, 0,} 3, 0, 0, 0)$ $(1, 0)$

Count fringe subtrees through the degree sequence

- Let $(\xi_1^n, \ldots, \xi_n^n)$ be the degree sequence of $\mathcal{T}_n^{gw}$.
- Let $(d_1, \ldots, d_{|T|})$ be the degree sequence of T.
- Then $N_T(\mathcal{T}_n^{gw})$ can be write as

$$
\begin{aligned}
N_T(\mathcal{T}_n^{gw}) &= \sum_{j=1}^{n} I_j \\
&\equiv \sum_{j=1}^{n} \mathbb{1}_{\left[(\xi_j^n, \ldots, \xi_{j+|T|-1}^n) = (d_1, \ldots, d_{|T|})\right]}.
\end{aligned}
$$

Why fringe subtrees are like unconditional Galton-Watson trees

- When n is large, $\xi_1^n, \ldots, \xi_n^n$ are close to $\xi_1, \ldots, \xi_n$ (n independent copies of ξ).
- Thus

$$\mathbb{P}\left\{I_j = 1\right\} = \mathbb{P}\left\{\cap_{i=1}^{|T|}\left[\xi_{j+i-1}^n = d_i\right]\right\}$$
$$\approx \prod_{i=1}^{|T|} \mathbb{P}\{\xi_i = d_i\} = \mathbb{P}\{\mathcal{T}^{gw} = T\} \equiv \pi(T).$$

- So $I_1, \ldots, I_n$ are close to iid Bernoulli $\pi(T)$.
- This is why $N_T(\mathcal{T}_n^{gw}) = \sum_{j=1}^n I_j \approx \mathrm{Bi}(n, \pi(T)) \approx \mathrm{Po}(n\pi(T))$.

The exchangeable pair method

- The proof of Theorem 1.2 uses the exchangeable pair method (Ross (2011, thm. 4.37)).
- It is a variation of Stein's method for Poisson distribution.

Example

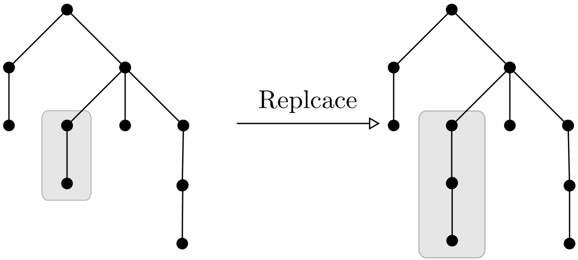
- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be iid $\mathrm{Be}(p)$.
- Let $W = X_1 + \cdots + X_n$.
- Let $W' = W - X_Z + Y_Z$ where $Z \stackrel{\mathcal{L}}{=} \mathrm{Unif}(\{1, \ldots, n\})$.
- We have an exchange pair — $(W, W') \stackrel{\mathcal{L}}{=} (W', W)$.
- Compute

$$\mathbb{P}\{W' = W - 1 \mid X_1, \ldots, X_n\}, \quad \mathbb{P}\{W' = W + 1 \mid X_1, \ldots, X_n\}.$$

- Then the method says $d_{TV}(W, \mathrm{Po}(\mathbb{E}W)) \leqslant p$.

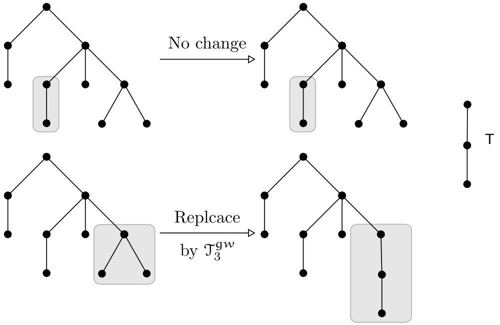
Subtree replacing – the naive way

- Recall $N_T(\mathcal{T}_n^{gw}) = \sum_{i=1}^{n} I_i$.
- What if we do the same thing for $N_T(\mathcal{T}_n^{gw})$?
- Let $\bar{N} = N_T(\mathcal{T}_n^{gw}) - I_Z + I'_Z$ with $I'_Z \stackrel{\mathcal{L}}{=} I_Z$.
- Is $(\bar{N}, N_T(\mathcal{T}_n^{gw}))$ an exchangeable pair?



Subtree replacing – the proper way

- Choose a fringe subtree of $\mathcal{T}_n^{gw}$ uniformly at random.
    - If its size is not the same as T, do nothing.
    - Otherwise, replace it with $\mathcal{T}_{|T|}^{gw}$.
- Let $\bar{N}$ be the number of T in the new tree.
- Then $(N_T(\mathcal{T}_n^{gw}), \bar{N})$ is an exchangeable pair.



Upper bound of the total variation distance

- Let $\mathfrak{T}_k$ be the set of all trees of size k.
- Let $\mathcal{S} \subseteq \mathfrak{T}_k$.
- Let $N_{\mathcal{S}}(\mathcal{T}_n^{gw})$ be the number of fringe subtrees that belongs to $\mathcal{S}$.
- Let $\pi(\mathcal{S}) \equiv \mathbb{P}\{\mathcal{T}^{gw} \in \mathcal{S}\}$.
- So $N_T(\mathcal{T}_n^{gw}) = N_{\{T\}}(\mathcal{T}_n^{gw})$.

Lemma 4.1

Let $k = k_n = o(n)$ *and* $k \to \infty$. *We have*

$$\sup_{\mathcal{S} \subseteq \mathfrak{T}_k} \frac{d_{TV}\left(N_{\mathcal{S}}(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(\mathcal{S}))\right)}{\pi(\mathcal{S})/\pi(\mathfrak{T}_k) + \sqrt{\pi(\mathcal{S})/\pi(\mathfrak{T}_k)}} \leqslant 1 + o\left(k^{-3/2}\right) + O\left(\frac{k^{1/4}}{\sqrt{n}}\right).$$

Large fringe subtrees count—set version

Theorem 1.3

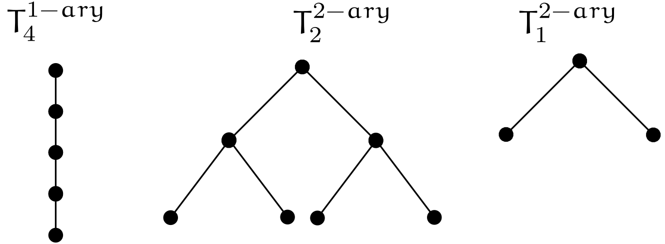
Let $\mathfrak{T}_k$ *be the set of trees of size* k*. Let* $k_n = o(n)$ *and* $k_n \to \infty$*. Let* $(\mathcal{S}_n)_{n \geqslant 1}$ *be a sequence with* $\mathcal{S}_n \subseteq \mathfrak{T}_{k_n}$*. We have:*

1. *If* $n\pi(\mathcal{S}_n) \to 0$*, then* $N_{\mathcal{S}_n}(\mathcal{T}_n^{gw}) = 0$ *whp.*
2. *If* $n\pi(\mathcal{S}_n) \to \mu \in (0, \infty)$*, then* $N_{\mathcal{S}_n}(\mathcal{T}_n^{gw}) \xrightarrow{d} \mathrm{Po}(\mu)$*.*
3. *If* $n\pi(\mathcal{S}_n) \to \infty$*, then*
$$\frac{N_{\mathcal{S}_n}(\mathcal{T}_n^{gw}) - n\pi(\mathcal{S}_n)}{\sqrt{n\pi(\mathcal{S}_n)}} \xrightarrow{d} N(0,1).$$
4. *If* $\pi(\mathcal{S}_n)/\pi(\mathfrak{T}_{k_n}) \to 0$*, then*
$$\lim_{n\to\infty} d_{TV}\left(N_{\mathcal{S}_n}(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(\mathcal{S}_n))\right) = 0.$$

Outline

Application 1—largest complete r-ary fringe subtree

- Let $T_h^{\text{r-ary}}$ be a complete r-ary tree of height h.



Application 1—largest complete r-ary fringe subtree

Lemma 5.2 & 5.3

Let $H_{n,r}$ *be the height of the largest complete* r*-ary fringe subtree in* $\mathcal{T}_n^{gw}$. *Then for* $r \geqslant 2$,

$$H_{n,r} - \log_r \log n \xrightarrow{p} -\alpha_r,$$

where α_r *is a constant. And*

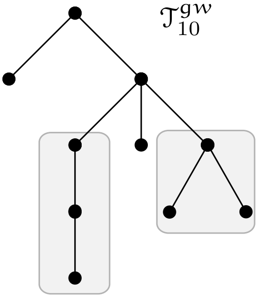
$$\frac{H_{n,1} \log(1/\mathbb{P}\{\xi = 1\})}{\log n} \xrightarrow{p} 1.$$

Method:

- Find the maximum h such that $n\pi(T_h^{r\text{-}ary}) \to \infty$.
- Then apply Theorem 1.2.

Application 2—existence of all possible subtrees

- Let K_n be the maximum k such that $\mathcal{T}_n^{gw}$ contains all trees of size $\leqslant k$ as fringe subtree.



$K_{10} = 3$

The coupon collector problem

Original version

There are n types of coupons. Each time we draw one type of coupon *uniformly at random*. How many draws do we need to collect all n types?

Generalized version

There are n types of coupons. Each time we draw a coupon, we get type i with probability p_i. How many draws do we need to collect all n types?

The coupon collector problem: the answer

Lemma 5.1 (Generalized coupon collector)

Assume X *takes values in* $\{1, \dots, n\}$. *Let* $p_i \equiv \mathbb{P}\{X = i\}$. *Let* $X_1, X_2, \dots$ *be i.i.d. copies of* X. *Let*

$$N \equiv \inf\{i \geqslant 1 : |\{X_1, X_2, \dots, X_i\}| = n\}.$$

Let m *be a positive integers. We have*

$$1 - \sum_{i=1}^{n} (1 - p_i)^m \leqslant \mathbb{P}\{N \leqslant m\} \leqslant \frac{1}{\sum_{i=1}^{n} (1 - p_i)^m}.$$

If $p_i = 1/n$, then $N = n\log(n) + o_p(1)$.

Connection to our problem

- Draw independent copies $\mathcal{T}_k^{gw}$ until every tree of size k has appeared.
- Let M_k be the number of draws.
- $N_{\mathfrak{T}_k}(\mathcal{T}_n^{gw}) \approx n\pi(\mathfrak{T}_k)$.
- So if $n\pi(\mathfrak{T}_k) > M_k$, then probably we have all trees of size k as fringe subtree, otherwise we do not.
- This is a coupon collector problem!

The least possible tree

- Among all coupons, there is one that is least likely to appear.
- If we get this one, we are likely to have all coupons.
- Let $T_k^{\min}$ be the least possible fringe subtree of size k.
- M_k depends on

$$p_k^{min} \equiv \mathbb{P}\left\{\mathcal{T}^{gw} = T_k^{\min}\right\}.$$

Lemma

- *If* $np_k^{min} \to 0$, *then* $T_k^{\min}$ *does not appear.*
- *If* $np_k^{min}/k \to \infty$, *then all possible subtrees of size* k *appear.*

What can we say about the least possible subtree?

- p_k^{min} certainly depends on ξ.
- But there is a small surprise.

Theorem 5.2

We have

$$(p_k^{min})^{1/k} \to L$$

as $k \to \infty$*, where* $0 \leqslant L < 1$ *is a constant defined as*

$$L \equiv \inf_{i \geqslant 1} \left\{ \mathbb{P}\{\xi = 0\} \left(\frac{\mathbb{P}\{\xi = i\}}{\mathbb{P}\{\xi = 0\}} \right)^{1/i} \right\}.$$

Threshold of existence of all possible subtrees

- By theorem 5.2, if $L > 0$, then $\log(1/p_k^{\min}) \sim k \log(1/L)$.
- $K_n = \log_{1/L} n + o_p(1)$ in this case.

Theorem 5.1

Assume that as $k \to \infty$,

$$\log(1/p_k^{\min}) \sim \gamma k^{\alpha} (\log k)^{\beta},$$

where $\alpha \geqslant 1$, $\beta \geqslant 0$, $\gamma > 0$ *are constants. Then*

$$\frac{K_n}{(\log n/(\log\log n)^{\beta})^{1/\alpha}} \xrightarrow{p} \left(\frac{\alpha^{\beta}}{\gamma}\right)^{1/\alpha}.$$

Applications

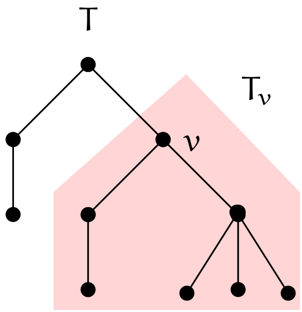
GW Tree	ξ	$\log(1/p_k^{min})$	K_n
Full binary trees	$2 \times \mathrm{Be}(1/2)$	$k \log 2$	$\log_2 n$
Motzkin trees	$\mathrm{Unif}(\{0,1,2\})$	$k \log 3$	$\log_3 n$
Binary trees	$\mathrm{Bi}(2,1/2)$	$k \log 4$	$\log_4 n$
d-ary trees	$\mathrm{Bi}(d,1/d)$	$k \log c_d$	$\log_{c_d} n$
Plane trees	$\mathrm{Ge}(1/2)$	$k \log 4$	$\log_4 n$
Cayley trees	$\mathrm{Po}(1)$	$k \log k$	$\frac{\log n}{\log \log n}$

- c_d is constant.
- Cayley tree is different because it has $L = 0$.

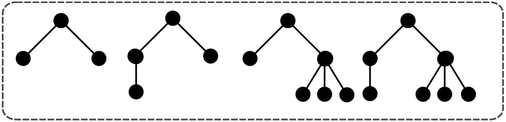
Outline

Non-fringe subtrees

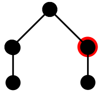
- Take a fringe subtree T_v.
- Replace some (or none) of T_v's own fringe subtrees with leaves.
- The result is a called a non-fringe subtree at v.



Non-fringe subtrees at v

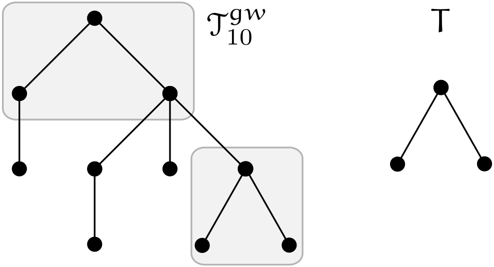


Not a non-fringe subtree !



Non-fringe subtree count

- Let $N_T^{nf}(\mathcal{T}_n^{gw})$ be the number of non-fringe subtrees of shape T in $\mathcal{T}_n^{gw}$.



$$N_T^{nf}(\mathcal{T}_{10}^{gw}) = 2$$

Large Non-fringe subtree Count

- Let $\pi^{nf}(T)$ be the prob. that $\mathcal{T}^{gw}$ has T as a non-fringe subtree at its root.
- We should have $N_T^{nf}(\mathcal{T}_n^{gw}) \approx \mathrm{Bi}(n, \pi^{nf}(T))$.

Theorem 1.4

Let T_n be a sequence of trees with $|T_n| = o(n)$. We have

1. *If $n\pi^{nf}(T_n) \to 0$, then $N_{T_n}^{nf}(\mathcal{T}_n^{gw}) = 0$ whp.*
2. *If $n\pi^{nf}(T_n) \to \infty$, then*

$$\frac{N_{T_n}^{nf}(\mathcal{T}_n^{gw})}{n\pi^{nf}(T_n)} \xrightarrow{p} 1.$$

Proof by computing first and second moments

Lemma 6.9 & 6.10

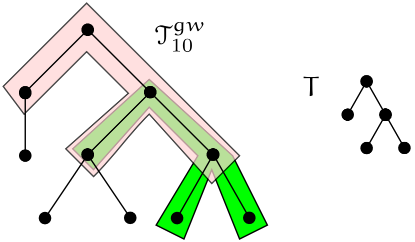
Assume that $|T_n| = o(n)$ *and* $n\pi^{nf}(T_n) \to \infty$. *We have*

1. $\mathbb{E}\left[N^{nf}_{T_n}(\mathcal{T}^{gw}_n)\right] = (1 + o(1))n\pi^{nf}(T_n)$.
2. $\mathrm{Var}\left(N^{nf}_{T_n}(\mathcal{T}^{gw}_n)\right) = o(n\pi^{nf}(T_n))^2$.

So Theorem 1.4 follows by Chebyshev's inequality.

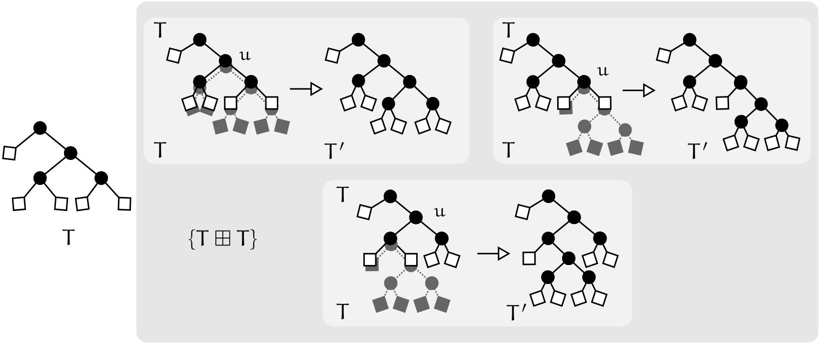
Difference between fringe and non-fringe subtrees

- Non-fringe subtrees can overlap.
- So it is more difficult to compute the second moment.



Glue two trees

Let $\{T \boxplus T\}$ be the trees that are two of T glued together.



The second factorial moment

Lemma 6.8

Assuming that $|T| = o(n)$, *we have*

$$\mathbb{E}\left[N_T^{nf}(\mathcal{T}_n^{gw})(N_T^{nf}(\mathcal{T}_n^{gw}) - 1)\right] \approx (n\pi^{nf}(T))^2 + 2n \sum_{T' \in \{T \boxplus T\}} \pi^{nf}(T')$$

- If the second term is $o(n\pi^{nf}(T))^2$ then we are done.
- Large $T' \in \{T \boxplus T\}$ should not be a problem.
- And there *cannot* be many small T' (with $|T'| < 3/2|T|$).

Application 1—largest complete r-ary non-fringe subtrees

Lemma 6.12 & 6.13

Let $\bar{H}_{n,r}$ *be the height of the largest complete* r*-ary non-fringe subtree in* $\mathcal{T}_n^{gw}$*. Then for* $r \geqslant 2$*,*

$$\bar{H}_{n,r} - \log_r \log n \xrightarrow{p} -\alpha'_r.$$

And

$$\frac{\bar{H}_{n,1} \log(1/\mathbb{P}\{\xi = 1\})}{\log n} \xrightarrow{p} 1.$$

Proof: Same as for fringe version.

Application 2—maximum degree

- A node of degree d can be seen as a non-fringe subtree that consists of the root and d-leaves.
- So Theorem 1.4 implies:

Theorem Meir and Moon (1991)

Assume that as $k \to \infty$,

$$\frac{1}{\mathbb{P}\{\xi = k\}^{1/k}} \to \rho > 1.$$

Let Y_n *be the maximum degree in* $\mathcal{T}_n^{gw}$, *then*

$$\frac{Y_n}{\log n} \xrightarrow{p} \frac{1}{\log \rho}.$$

Open questions

- For fringe subtrees, does

$$d_{TV}\left(N_{\mathfrak{T}_k}(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(\mathfrak{T}_k)\right) \to 0,$$

  as $k \to \infty$?
- For non-fringe subtrees
  - A central limit theorem?
  - What is the total number of non-fringe subtrees in $\mathcal{T}_n^{gw}$?

Bibliography

D. Aldous, "Asymptotic fringe distributions for general families of random trees," *The Annals of Applied Probability*, vol. 1, no. 2, pp. 228–266, 1991.

I. J. Bienaymé, "De la loi de multiplication et de la durée des familles," *Société Philomatique Paris*, 1845, Reprinted in Kendall (1975).

X. S. Cai, "A study of large fringe and non-fringe subtrees in conditional Galton-Watson trees," PhD thesis, McGill University, Aug. 2016.

X. S. Cai and L. Devroye, "A study of large fringe and non-fringe subtrees in conditional Galton-Watson trees," *Latin American Journal of Probability and Mathematical Statistics*, 2017, To appear. arXiv: `1602.03850 [math.PR]`.

S. Janson, "Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton-Watson trees," *Random Structures and Algorithms*, vol. 48, no. 1, pp. 57–101, 2016.

A. Meir and J. W. Moon, "On nodes of large out-degree in random trees," *Congressus Numerantium*, vol. 82, pp. 3–13, 1991.

N. Ross, "Fundamentals of Stein's method," *Probability Surveys*, vol. 8, pp. 210–293, 2011.

My coauthor