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

The Discrete Math Seminar

Xing Shi Cai, Luc Devroye

Feb 25, 2022

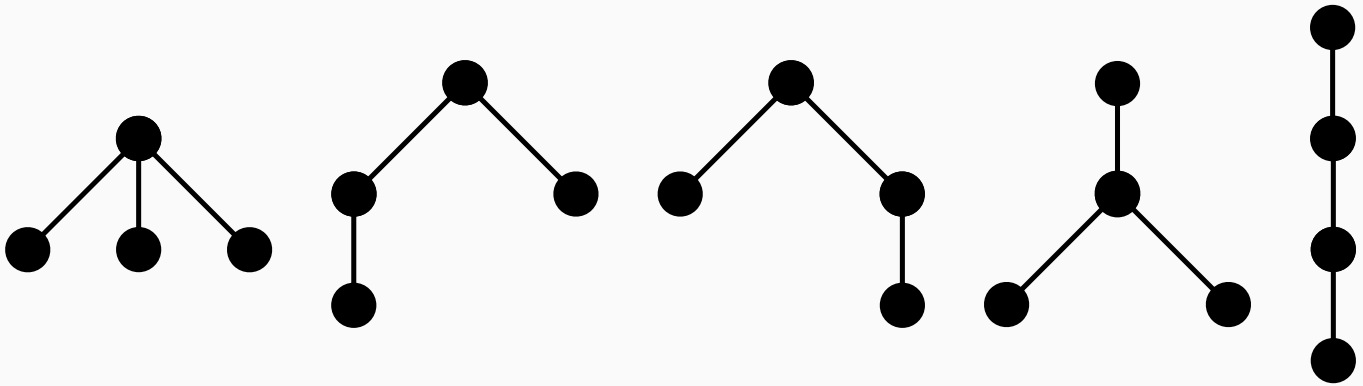
Duke Kunshan University

Introduction

What is a tree

A tree is an acyclic graph.

In this talk, trees are *unlabelled*, *rooted*, and *ordered* (plane trees).

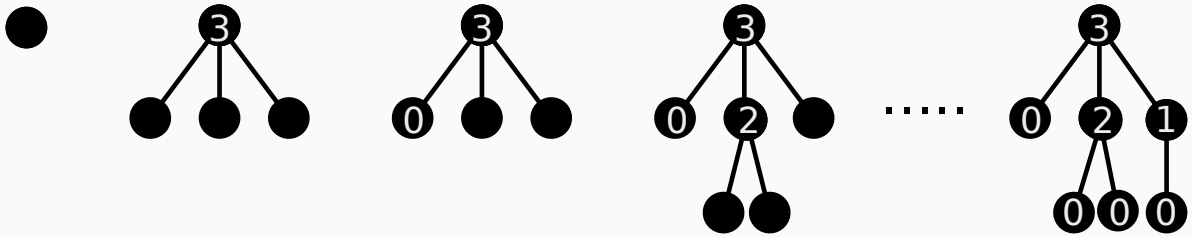


Galton-Watson trees

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 (offspring distribution).

Introduced by **bienayme**.



Note

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

Conditional Galton-Watson trees

A conditional/conditioned GW Tree $\mathcal{T}_n^{gw}$ is $\mathcal{T}^{gw}$ restricted to $|\mathcal{T}^{gw}| = n$.

So $\mathbb{P}\left\{\mathcal{T}_n^{gw} = T\right\} = \mathbb{P}\left\{\mathcal{T}^{gw} = T \mid |\mathcal{T}^{gw}| = n\right\}$.

It covers many uniform random tree models (Janson, 2012) –

- full binary trees
- binary trees
- d-ary trees
- Motzkin trees
- Plane trees
- Cayley trees

Simply generated trees

In many cases, $\mathcal{T}_n^{gw}$ is equivalent to simply generated trees introduced by Meir and Moon (1978).

Let $(w_i)_{i \geq 0}$ be a sequence of non-negative numbers.

Let $\mathbf{weight}(T) = \prod_{v \in T} w_{\mathbf{deg}(v)}$.

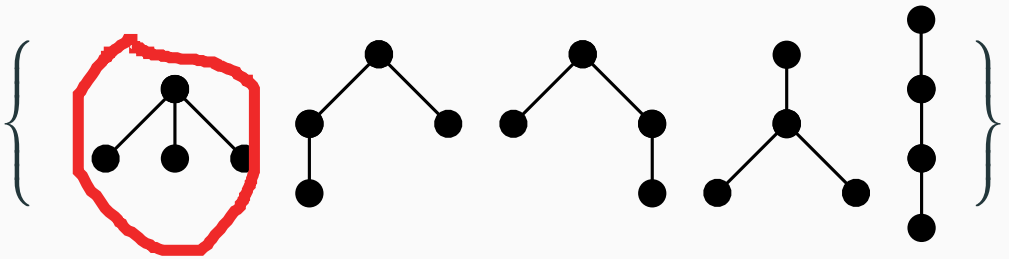
Let $\mathcal{T}_n^{\mathrm{sg}}$ be a random tree of size n such that $\mathbb{P}\left\{\mathcal{T}_n^{\mathrm{sg}} = T\right\}$ is proportional to $\mathbf{weight}(T)$.

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.

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

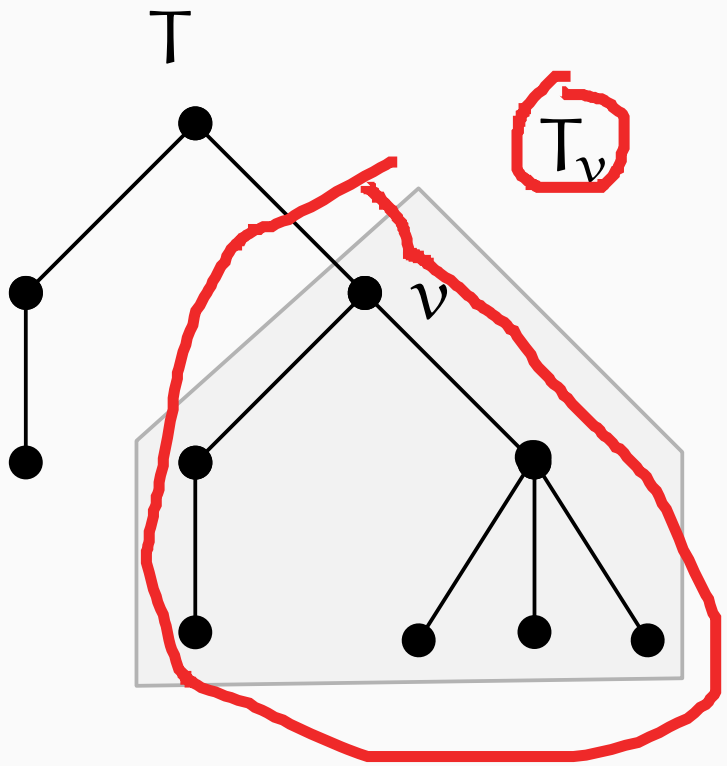


trees of size n = 4

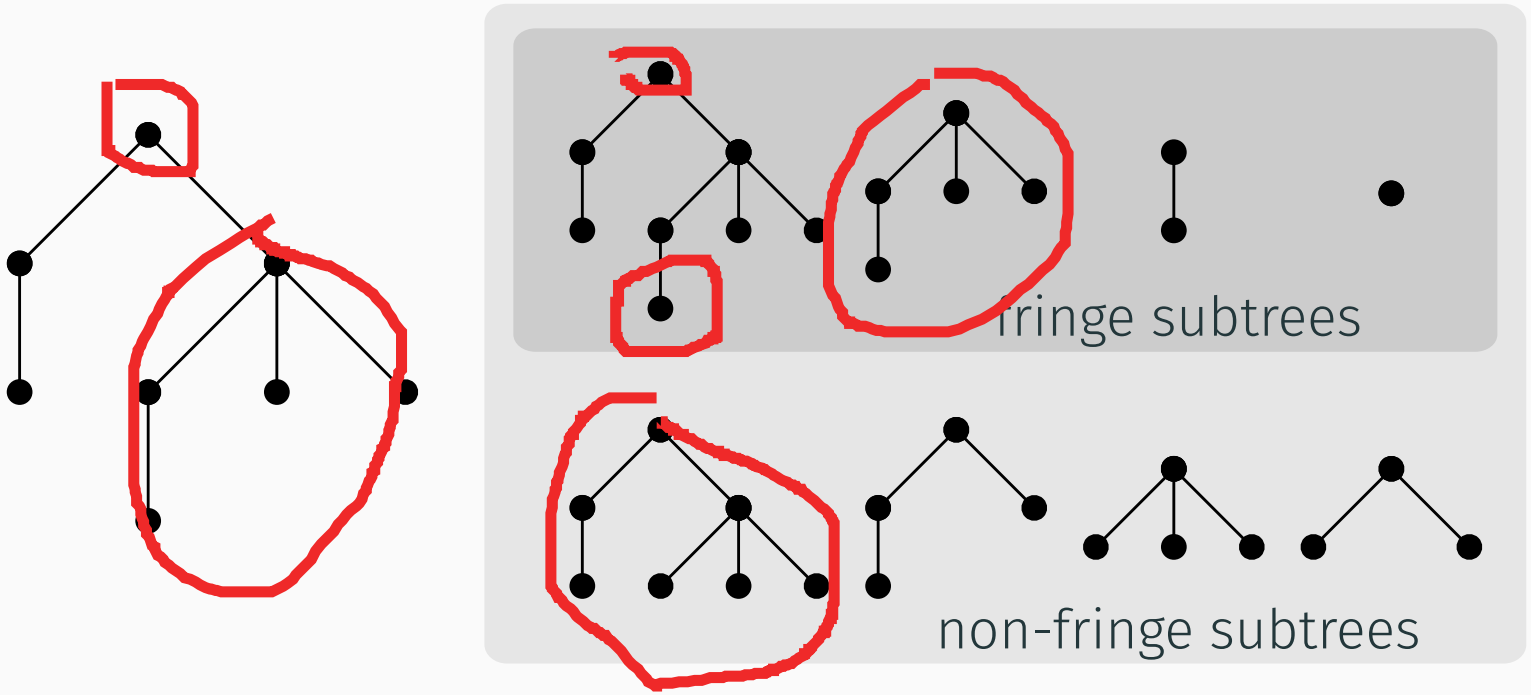
Fringe subtrees

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 and non-fringe subtrees

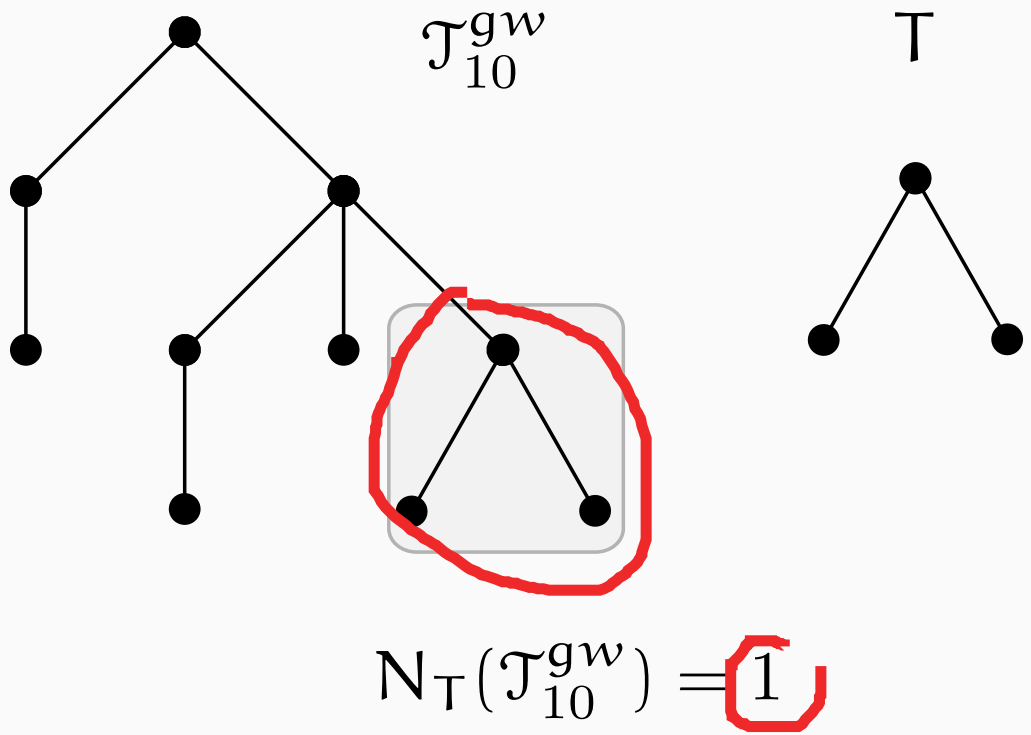


Fringe subtree—a node and every descendant of it.

Non-fringe subtree—a fringe subtree with some (or none) of its subtrees “trimmed”—replaced by leaves.

Fringe subtree count

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

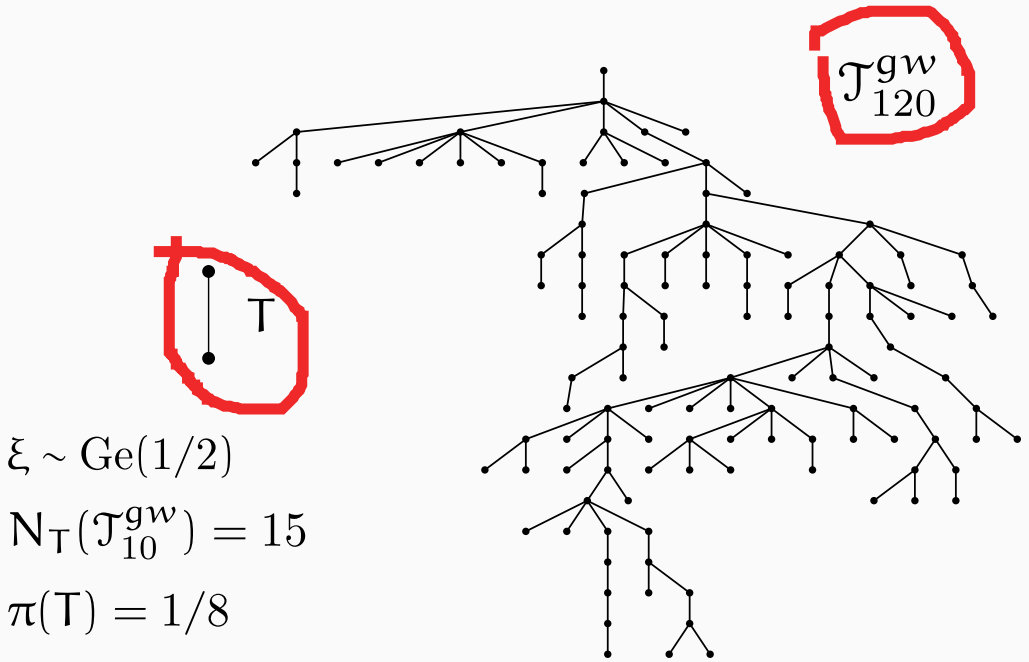


Fringe subtree count: bigger example

In the next example,

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

Is this just a coincidence?



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

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

$\pi(T) = 1/8$

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

Law of Large Number (Aldous, 1991)

As $n \to \infty$,

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

Central Limit Theorem (Janson, 2016)

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 $\leq k$ as fringe subtree.
- What about non-fringe subtrees?

Large Fringe Subtrees

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 (Cai, 2016)

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

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

Large fringe subtrees cont.

Theorem 1.2 (Cai, 2016)

So letting $(T_n)_{n\geq 1}$ be a sequence of trees with $|T_n| = k_n$,

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

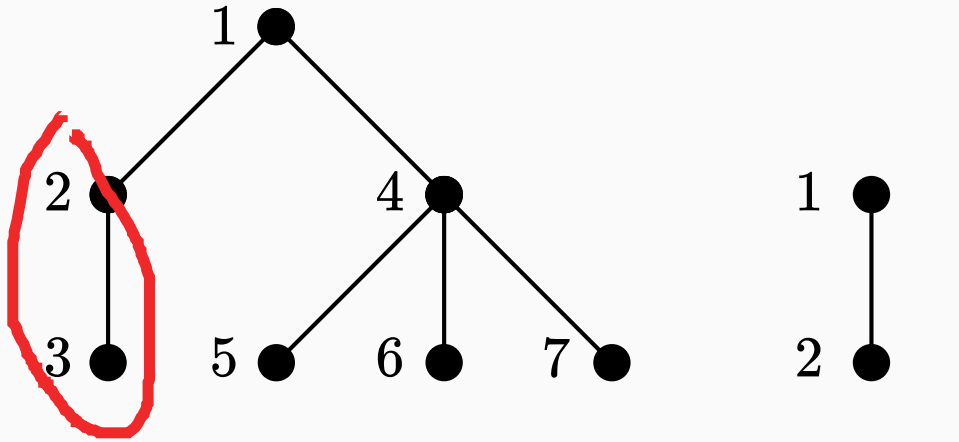
$$\frac{N_{T_n}(\mathcal{T}_n^{gw}) - n\pi(T_n)}{\sqrt{n\pi(T_n)}} \xrightarrow{d} N(0,1).$$

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 (DFS) order.

We can count fringe subtree through degree sequence.



Degree sequence: $(2, 1, 0, 3, 0, 0, 0)$ $(1, 0)$

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

$$N_T(\mathcal{T}_n^{gw}) = \sum_{i=1}^{n} I_i$$

$$\equiv \sum_{i=1}^{n} \mathbb{1}_{\left[(\xi_i^n, \ldots, \xi_{i+|T|-1}^n) = (d_1, \ldots, d_{|T|})\right]}.$$

Why fringe subtrees are like unconditional GW 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

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

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 uses the exchangeable pair method (Ross, 2011, thm. 4.37) – 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 $Z \overset{\mathcal{L}}{=} \mathrm{Unif}(\{1, \ldots, n\})$.
- We have an exchange pair – $(W, W') \overset{\mathcal{L}}{=} (W', W)$.
- Compute

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

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

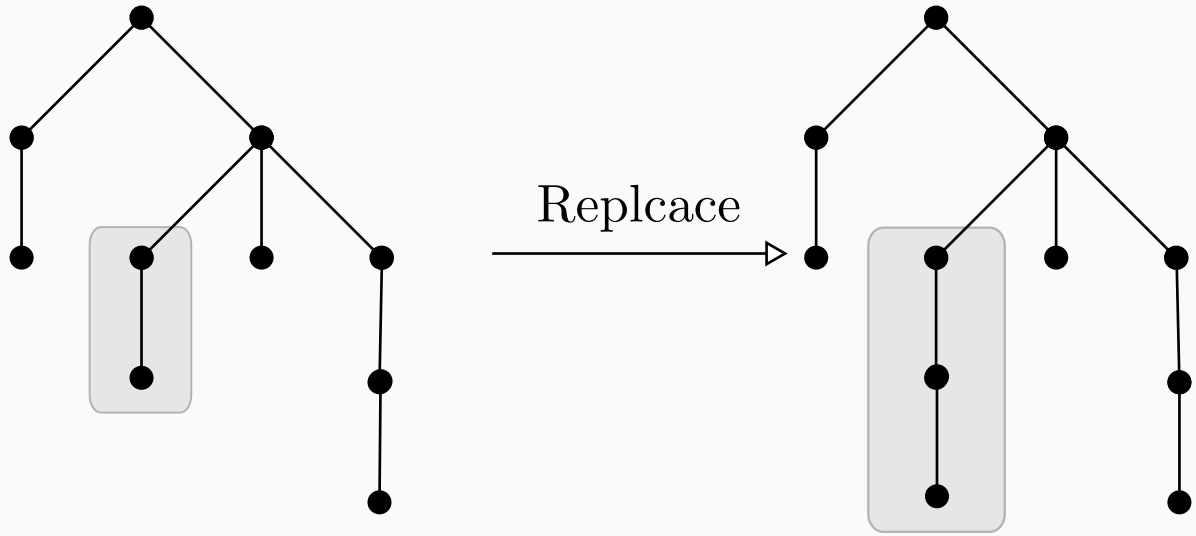
Subtree replacing – the naive way

Recall $N_T(\mathcal{T}_n^{gw}) = \sum_{i=1}^{n} l_i$.

What if we do the same thing for $N_T(\mathcal{T}_n^{gw})$?

Let $\bar{N} = N_T(\mathcal{T}_n^{gw}) - l_Z + l'_Z$ with $l'_Z \overset{\mathcal{L}}{=} l_Z$.

Is $(\bar{N}, N_T(\mathcal{T}_n^{gw}))$ an exchangeable pair?



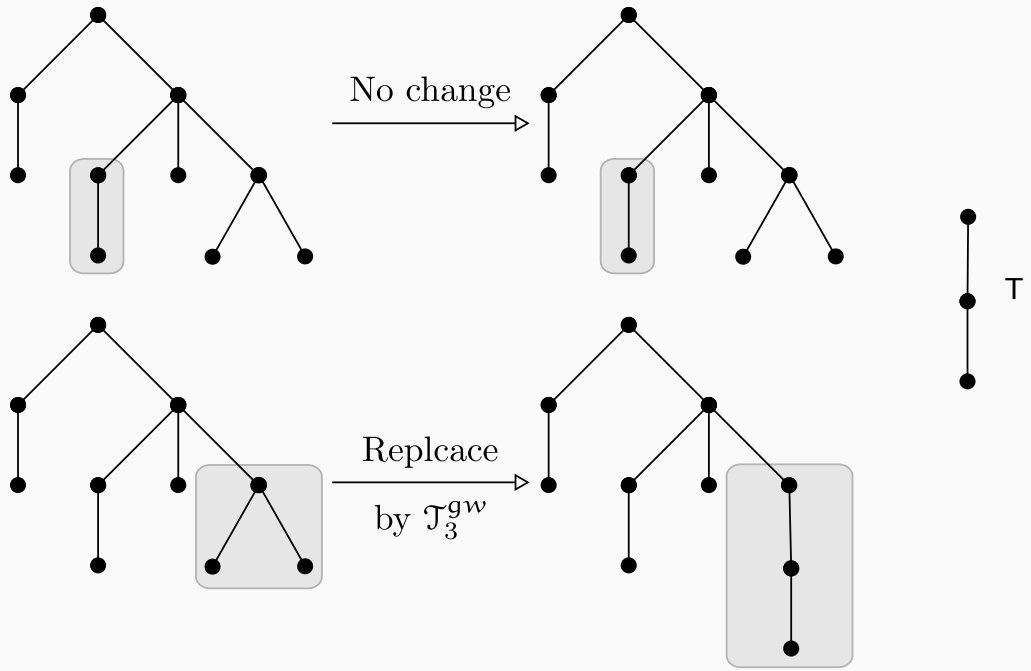
Subtree replacing – the proper way

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.



A generalization

Let $\mathfrak{T}_k$ be the set of all trees of size 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 $N_T(\mathcal{T}_n^{gw}) = N_{\{T\}}(\mathcal{T}_n^{gw})$.

Theorem 4.1

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

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

Idea of proof

There are roughly $n\pi(\mathfrak{T}_k)$ fringe subtrees of size k.

Each of them is itself a conditional GW tree of size k.

So $N_{\mathcal{S}}(\mathcal{T}_n^{gw}) \approx \mathrm{Bi}(n\pi(\mathfrak{T}_k), \pi(\mathcal{S})/\pi(\mathfrak{T}_k))$.

We know that

$$d_{\mathrm{TV}}\left(\mathrm{Bi}(m,p), \mathrm{Po}(mp)\right) \le p.$$

Then we should have

$$d_{\mathrm{TV}}\left(N_{\mathcal{S}}(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(\mathcal{S}))\right) \le \frac{\pi(\mathcal{S})}{\pi(\mathfrak{T}_k)}.$$

Large fringe subtrees count—set version

Theorem 1.3 (Cai, 2016)

Let $\mathfrak{T}_k$ be the set of trees of size k. Let $k_n = o(n)$ and $k_n \to \infty$. Let $(\mathcal{S}_n)_{n\geq 1}$ 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

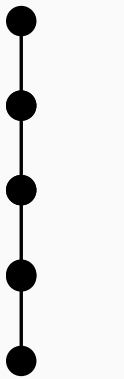
$$\lim_{n\to\infty} d_{\mathrm{TV}}\left(N_{\mathcal{S}_n}(\mathcal{T}_n^{gw}), \mathrm{Po}(n\pi(\mathcal{S}_n))\right) = 0.$$

Large Fringe Subtrees—Applications

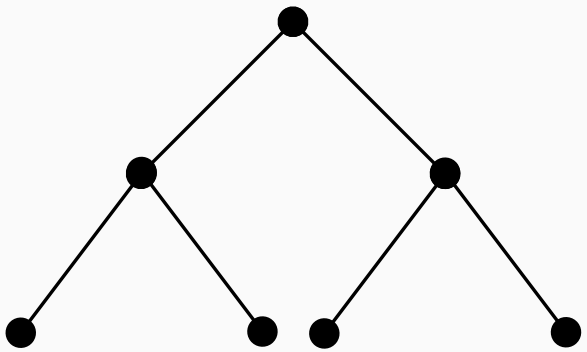
Application 1—largest complete r-ary fringe subtree

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

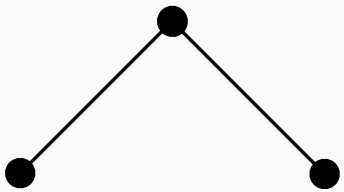
$T_4^{1-\mathrm{ary}}$



$T_2^{2-\mathrm{ary}}$



$T_1^{2-\mathrm{ary}}$



Application 1—largest complete r-ary fringe subtree

Theorem 5.2 & 5.3 (Cai, 2016)

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

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

where α_r is a constant. And

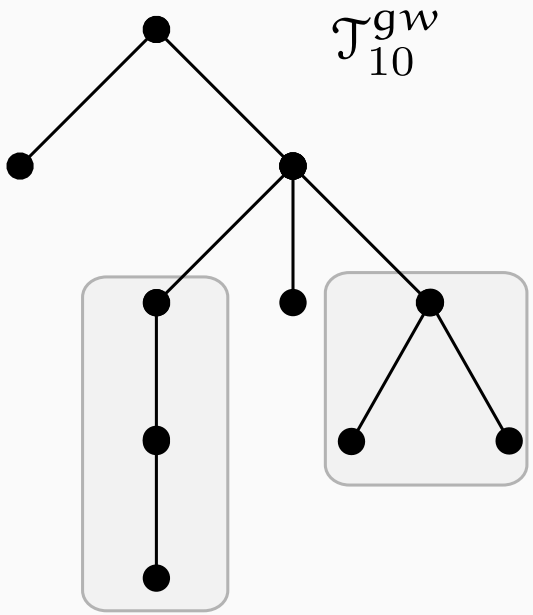
$$\frac{H_{n,1} \log(1/\mathbb{P}\{\xi = 1\})}{\log n} \xrightarrow{p} 1.$$

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 $\leq k$ as fringe subtree.



$$\mathrm{K}_{10} = 3$$

The coupon collector problem

Original version

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

Generalized version

There are n different 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 types of coupon?

The coupon collector problem: the answer

Theorem 5.1 (Generalized coupon collector) (Cai, 2016)

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

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

Let m be a positive integers. We have

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

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 $K_n \geq k$, otherwise $K_n < k$.
- This is a coupon collector problem!

Large Non-Fringe Subtrees

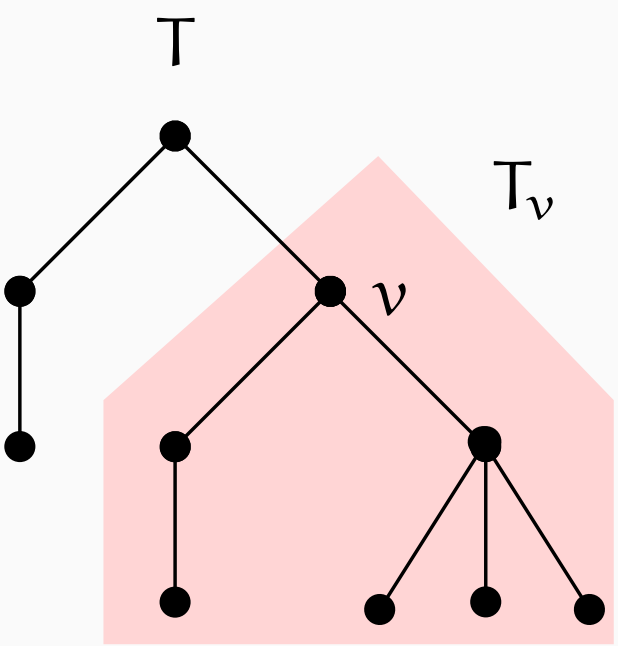
Non-fringe subtrees

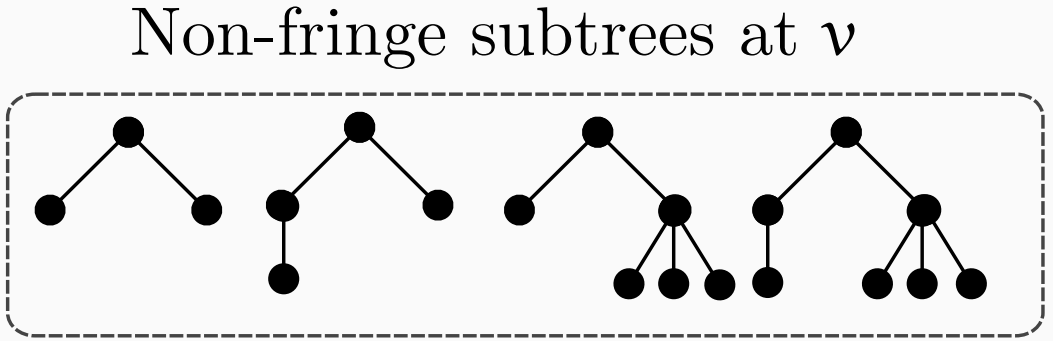
Take a fringe subtree T_v.

Replace some (or none) of T_v's own fringe subtrees with leaves.

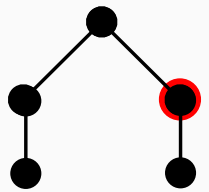
The result is a called a non-fringe subtree at v.

T_v is also a non-fringe subtree.



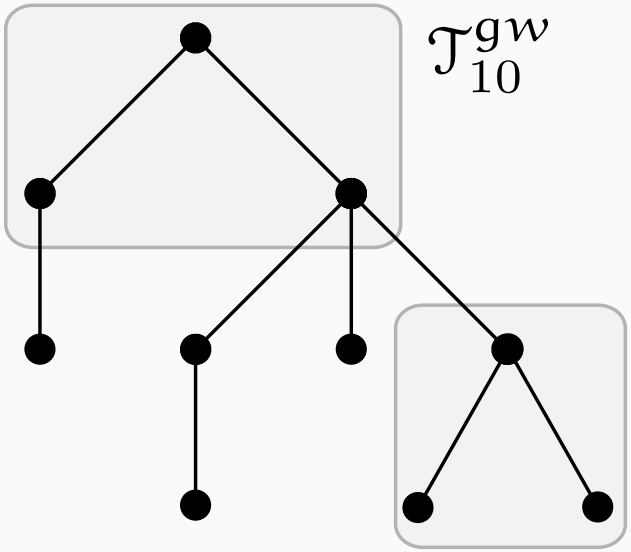


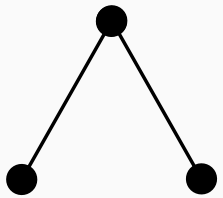
Not a non-fringe subtree !



Non-fringe subtree count

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





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

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 (Cai, 2016)

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

Theorem 6.9 & 6.10 (Cai, 2016)

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

1. $\mathbb{E}\left[N_{T_n}^{nf}(\mathcal{T}_n^{gw})\right] = (1+o(1))n\pi^{nf}(T_n)$.
2. $\operatorname{var} N_{T_n}^{nf}(\mathcal{T}_n^{gw}) = o(n\pi^{nf}(T_n))^2$.

So Theorem 1.4 follows by Chebyshev's inequality.

References i

[1] D. Aldous, "Asymptotic Fringe Distributions for General Families of Random Trees," *The Annals of Applied Probability*, vol. 1, no. 2, pp. 228–266, 1991.

[2] X. S. Cai, "A study of large fringe and non-fringe subtrees in conditional galton-watson trees," Ph.D. dissertation, 2016.

[3] X. S. Cai and L. Devroye, "A study of large fringe and non-fringe subtrees in conditional Galton-Watson trees," *ALEA Lat. Am. J. Probab. Math. Stat.*, vol. 14, no. 1, pp. 579–611, 2017.

References ii

[4] S. Janson, “Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation,” *Probab. Surv.*, vol. 9, pp. 103–252, 2012.

[5] S. Janson, “Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton–Watson trees,” *Random Structures & Algorithms*, vol. 48, no. 1, pp. 57–101, 2016.

[6] A. Meir and J. W. Moon, “On the altitude of nodes in random trees,” *Canad. J. Math.*, vol. 30, no. 5, pp. 997–1015, 1978.

[7] N. Ross, “Fundamentals of Stein’s method,” *Probability Surveys*, vol. 8, no. none, pp. 210–293, 2011.