Asymptotic normality of fringe subtrees and additive functionals in conditioned Galton–Watson trees

Svante Janson

AofA, Paris, June 2014

Fringe subtrees

A *fringe subtree* in a rooted tree T is a subtree consisting of a node v and all its descendents. We denote this tree by T_v.

The *random fringe subtree* T_* is the random rooted tree obtained by taking the subtree T_v at a uniformly random node v in T, see Aldous (1991).

Subtree counts

Let $\mathfrak{T}$ be the set of all finite rooted ordered trees. Let, for rooted trees $T, T' \in \mathfrak{T}$,

$$n_{T'}(T) := \big|\{v \in T : T_v = T'\}\big|,$$

i.e., the number of subtrees of T that are equal (isomorphic) to T'. Then the distribution of T_* is given by

$$\mathbb{P}(T_* = T') = n_{T'}(T)/|T|, \qquad T' \in \mathfrak{T}.$$

Thus, to study the distribution of T_* is equivalent to studying the numbers $n_{T'}(T)$.

Additive functionals

Let f be a functional of rooted trees, i.e., a function $f : \mathfrak{T} \to \mathbb{R}$, and for a tree $T \in \mathfrak{T}$ consider the sum

$$F(T) = F(T; f) := \sum_{v \in T} f(T_v).$$

Thus,

$$F(T)/|T| = \mathbb{E}\, f(T_*).$$

One important example of this is to take $f(T) = \mathbf{1}\{T = T'\}$, the indicator function that T equals some given tree $T' \in \mathfrak{T}$; then $F(T) = n_{T'}(T)$. Conversely, for any f,

$$F(T) = \sum_{T' \in \mathfrak{T}} f(T') n_{T'}(T);$$

hence any $F(T)$ can be written as a linear combination of the subtree counts $n_{T'}(T)$, so the two points of views are equivalent.

Functionals F of this type are called *additive functionals*. The definition above can also be written recusively as

$$F(T) = f(T) + \sum_{i=1}^{d} F(T_i),$$

where $T_1, \ldots, T_d$ are the branches (i.e., the subtrees rooted at the children of the root) of T.

$f(T)$ is often called a *toll function*.

In our case, T is a random tree, and then $F(T)$ is a random variable. In particular, $n_{T'}(T)$ is a random variable for each $T' \in \mathfrak{T}$, and thus the distribution of T_*, is a random probability distribution on $\mathfrak{T}$, with

$$\mathbb{P}\big(T_* = T' \mid T\big) = n_{T'}(T)/|T|$$

Similarly

$$F(T)/|T| = \mathbb{E}\big(f(T_*) \mid T\big).$$

Galton-Watson trees

The random trees that we consider are conditioned Galton–Watson trees. (Related results for some other random trees are given by Fill and Kapur (*m*-ary search trees under different models) and Holmgren and Janson (random binary search trees and random recursive trees).)

The Galton–Watson trees are defined using an offspring distribution ξ and we assume that $\mathbb{E}\,\xi = 1$ and $\sigma^2 := \mathrm{Var}\,\xi$ is finite (and non-zero). Let $p_k := \mathbb{P}(\xi = k)$.

Law of large numbers

Theorem (Aldous, et al.)

Let $\mathcal{T}_n$ be a conditioned Galton–Watson tree with n nodes, defined by an offspring distribution ξ with $\mathbb{E}\,\xi = 1$, and let $\mathcal{T}$ be the corresponding unconditioned Galton–Watson tree. Then, as $n \to \infty$: For every fixed tree T,

$$\frac{n_T(\mathcal{T}_n)}{n} = \mathbb{P}(\mathcal{T}_{n,*} = T \mid \mathcal{T}_n) \xrightarrow{\text{p}} \mathbb{P}(\mathcal{T} = T).$$

Equivalently, for any bounded functional f on $\mathfrak{T}$,

$$\frac{F(\mathcal{T}_n)}{n} = \mathbb{E}\,f\left(\mathcal{T}_{n,*} \mid \mathcal{T}_n\right) \xrightarrow{\text{p}} \mathbb{E}\,f(\mathcal{T}).$$

A central limit theorem

Theorem

Let $\mathcal{T}_n$ be a conditioned Galton–Watson tree of order n with offspring distribution ξ, where $\mathbb{E}\,\xi = 1$ and $0 < \sigma^2 := \operatorname{Var}\xi < \infty$, and let $\mathcal{T}$ be the corresponding unconditioned Galton–Watson tree. Suppose that $f : \mathfrak{T} \to \mathbb{R}$ is a functional of rooted trees such that $\mathbb{E}\,|f(\mathcal{T})| < \infty$, and let $\mu := \mathbb{E}\,f(\mathcal{T})$.

(i) *If $\mathbb{E}\,f(\mathcal{T}_n) \to 0$ as $n \to \infty$, then*

$$\mathbb{E}\,F(\mathcal{T}_n) = n\mu + o\big(\sqrt{n}\big).$$

Theorem, cont.

(ii) If

$$\mathbb{E}\, f(\mathcal{T}_n)^2 \to 0$$

as $n \to \infty$, and

$$\sum_{n=1}^{\infty} \frac{\sqrt{\mathbb{E}(f(\mathcal{T}_n)^2)}}{n} < \infty,$$

then

$$\operatorname{Var} F(\mathcal{T}_n) = n\gamma^2 + o(n)$$

where

$$\gamma^2 := 2\,\mathbb{E}\Big(f(\mathcal{T})\big(F(\mathcal{T}) - |\mathcal{T}|\mu\big)\Big) - \operatorname{Var} f(\mathcal{T}) - \mu^2/\sigma^2$$

is finite; moreover,

$$\frac{F(\mathcal{T}_n) - n\mu}{\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, \gamma^2).$$

Remarks

Special cases of the theorem have been proved before, by various methods. A simple example is the number of leaves in $\mathcal{T}_n$, shown to be normal by Kolchin (1986).

Wagner (2012) considered random labelled trees (the case $\xi \sim \text{Po}(1)$) and showed the result above (and convergence of all moments) under stronger hypotheses on f.

Joint convergence for several different F (each satisfying the conditions in the theorem) follows immediately by the Cramér–Wold device. For example:

Corollary

The subtree counts $n_T(\mathcal{T}_n)$, $T \in \mathfrak{T}$, are asymptotically jointly normal. More precisely, let $\pi_T := \mathbb{P}(\mathcal{T} = T)$,

$$\gamma_{T,T} := \pi_T - \bigl(2|T| - 1 + \sigma^{-2}\bigr)\pi_T^2,$$

and, for $T_1 \neq T_2$,

$$\gamma_{T_1,T_2} := n_{T_2}(T_1)\pi_{T_1} + n_{T_1}(T_2)\pi_{T_2} - \bigl(|T_1| + |T_2| - 1 + \sigma^{-2}\bigr)\pi_{T_1}\pi_{T_2}.$$

Then, for any trees $T, T_1, T_2 \in \mathfrak{T}$,

$$\begin{aligned} \mathbb{E}\, n_T(\mathcal{T}_n) &= n\pi_T + o\bigl(\sqrt{n}\bigr), \\ \operatorname{Cov}\bigl(n_{T_1}(\mathcal{T}_n), n_{T_2}(\mathcal{T}_n)\bigr) &= n\gamma_{T_1,T_2} + o(n), \\ \frac{n_T(\mathcal{T}_n) - n\pi_T}{\sqrt{n}} &\xrightarrow{\mathrm{d}} Z_T, \end{aligned}$$

jointly for all $T \in \mathfrak{T}$, *where* Z_T *are jointly normal with mean* $\mathbb{E}\, Z_T = 0$ *and covariances* $\operatorname{Cov}\bigl(Z_{T_1}, Z_{T_2}\bigr) = \gamma_{T_1,T_2}$.

The conditions on f say that $f(T)$ is (on the average, at least) decreasing as $|T| \to \infty$, but a rather slow decrease is sufficient; for example, the theorem applies when $f(T) = 1/\log^2 |T|$.

It is in general *not* enough to assume that f is a bounded functional. However, the following holds.

Theorem

The central limit theorem extends to all bounded functionals f that are local, i.e. depend only on the first M generations of T for some fixed M.

Remark

For $f(T)$ that grow with the size $|T|$, we cannot expect the results to hold. See Fill and Kapur (2004) for the case of binary trees. They show that for $f(T) = \log|T|$, $F(\mathcal{T}_n)$ is asymptotically normal, but with a variance of the order $n \log n$. And if $f(T) = |T|^\alpha$ for some $\alpha > 0$, then the variance is of order $n^{1+2\alpha}$, and $F(\mathcal{T}_n)$ has, after normalization, a non-normal limiting distribution.

Remark

For the m-ary search tree ($2 \le m \le 26$) and random recursive tree a similar theorem holds, but there $f(T)$ may grow almost as $|T|^{1/2}$, see Hwang and Neininger (2002) (binary search tree, f depends on $|T|$ only), Fill and Kapur (2005) (m-ary search tree, f depends on $|T|$ only), Holmgren and Janson (AofA 2014) (binary search tree and random recursive tree, general f). A reason for this difference is that for a conditioned Galton–Watson tree, the limit distribution of the size of the fringe subtree, which is the distribution of $|\mathcal{T}|$, decays rather slowly, with $\mathbb{P}(|\mathcal{T}| = n) \asymp n^{-3/2}$. while the corresponding limit distribution for fringe subtrees in a binary search tree or random recursive tree decays somewhat faster, as n^{-2}.

Problem

The asymptotic variance γ^2 equals 0 in two trivial cases:
(i) $f(\mathcal{T}) = F(\mathcal{T}) = F(\mathcal{T}_n) = 0$ a.s.;
(ii) $\{k : p_k > 0\} = \{0, r\}$ for some $r > 1$ and $f(\mathcal{T}) = a\mathbf{1}\{|\mathcal{T}| = 1\}$ for some real a; then $F(\mathcal{T}_n) = a(n - (n-1)/r)$ is deterministic. (The tree is r-ary and F is a times the number of leaves.)
We can show that if f has finite support, then $\gamma^2 > 0$ except in these trivial cases. For general f, we do not know whether $\gamma^2 = 0$ is possible except in these and related trivial cases.

Is $\gamma^2 = 0$ possible except when $F(\mathcal{T}_n)$ is deterministic for every n?

Examples

Example

$f(T) = \mathbf{1}\{|T| = 1\}$. Then $F(T)$ is the number of leaves in T. We have $\mathbb{E}\, f(\mathcal{T}) = \mathbb{P}(|\mathcal{T}| = 1) = \mathbb{P}(\xi = 0) = p_0$.
The theorem yields asymptotic normality with

$$\gamma^2 = 2p_0(1 - p_0) - p_0(1 - p_0) - p_0^2/\sigma^2 = p_0 - (1 + \sigma^{-2})p_0^2.$$

The asymptotic normality in this case (and a local limit theorem) was proved by Kolchin (1986).

Example

Let $n_r(T)$ be the number of nodes of outdegree r. Then $n_r(T) = F(T)$ with $f(T) = 1$ if the root of T has degree r, and $f(T) = 0$ otherwise. Asymptotic normality of $n_r(\mathcal{T}_n)$ too was proved by Kolchin (1986), see also Janson (2001) (joint convergence and moment convergence, assuming at least $\mathbb{E}\,\xi^3 < \infty$), Minami (2005) and Drmota's book (2009) (both assuming an exponential moment) for different proofs. Similarly, we obtain joint convergence for different r. (It seems that joint convergence has not been proved before without assuming at least $\mathbb{E}\,\xi^3 < \infty$.)

In this example, f does not decrease and the main theorm does not apply, but the version for bounded local f does.

Nevertheless, this result is a bit disappointing, since we do not obtain the Kolchin's explicit formula

$$\gamma_r^2 = p_r(1 - p_r) - (r-1)^2 p_r^2/\sigma^2$$

for the variance. The theorem shows existence of γ^2 but the formula given by the proof is rather involved, and we do not know any way to derive the explicit value from it.
(In this example, a special argument works.)

Example

A node in a (rooted) tree is said to be *protected* if it is neither a leaf nor the parent of a leaf. Convergence in probability of the fraction of protected nodes is proved for general conditioned Galton–Watson trees by Devroye and Janson (2013).

We can extend this to asymptotic normality. We define $f(T) := \mathbf{1}\{\text{the root of } T \text{ is protected}\}$, and then $F(T)$ is the number of protected nodes in T. Again, the version for bounded and local f applies.

The asymptotic mean $\mu = \mathbb{E}\, f(\mathcal{T})$ is easily calculated, see Devroye and Janson (2013).

However, we do not see how to find an explicit value of γ^2.

Example

Wagner (2012) studied the number $s(T)$ of arbitrary subtrees (not necessarily fringe subtrees) of the tree T, and the number $s_1(T)$ of such subtrees that contain the root.

He noted that if T has branches $T_1, \dots, T_d$, then $s_1(T) = \prod_{i=1}^{d}(1 + s_1(T_i))$ and thus

$$\log\bigl(1 + s_1(T)\bigr) = \log\bigl(1 + s_1(T)^{-1}\bigr) + \sum_{i=1}^{d} \log\bigl(1 + s_1(T_i)\bigr),$$

so $\log\bigl(1 + s_1(T)\bigr)$ is an additive functional with toll function $f(T) = \log\bigl(1 + s_1(T)^{-1}\bigr)$. He used this to show asymptotic normality of $\log s_1(\mathcal{T}_n)$ and $\log s(\mathcal{T}_n)$ for for the case of uniform random labelled trees.

We can generalize this to arbitrary conditioned Galton–Watson trees with $\mathbb{E}\,\xi = 1$ and $\mathbb{E}\,\xi^2 < \infty$.

Sketch of proofs

Let $\xi_1, \xi_2, \ldots$ be a sequence of independent copies of ξ, and let

$$S_n := \sum_{i=1}^{n} \xi_i.$$

A tree is uniquely described by its degree sequence $(d_1, \dots, d_n)$. We may thus define the functional f also on finite nonnegative integer sequences $(d_1, \dots, d_n)$, $n \geq 1$, by

$$f(d_1, \dots, d_n) := \begin{cases} f(T), & (d_1, \dots, d_n) \text{ is the degree sequence of a tree } T \\ 0, & \text{otherwise.} \end{cases}$$

If T has degree sequence $(d_1, \dots, d_n)$, and its nodes are numbered in depth-first order, then the subtree T_{v_i} has degree sequence $(d_i, d_{i+1}, \dots, d_{i+k-1})$, where $k \leq n - i + 1$ is the unique index such that $(d_i, \dots, d_{i+k-1})$ is a degree sequence of a tree. Thus,

$$F(T) = \sum_{1 \leq i \leq j \leq n} f(d_i, \dots, d_j) = \sum_{k=1}^{n} \sum_{i=1}^{n-k+1} f(d_i, \dots, d_{i+k-1}).$$

Moreover, if we regard $(d_1, \dots, d_n)$ as a cyclic sequence and define $d_{n+i} := d_i$, also

$$F(T) = \sum_{k=1}^{n} \sum_{i=1}^{n} f(d_i, \dots, d_{i+k-1}).$$

It well-known that up to a cyclic shift, the degree sequence $(d_1, \dots, d_n)$ of the conditioned Galton–Watson tree $\mathcal{T}_n$ has the same distribution as $\big((\xi_1, \dots, \xi_n) \mid S_n = n-1\big)$. Since the final sum above is invariant under cyclic shifts of $(d_1, \dots, d_n)$, it follows that

$$F(\mathcal{T}_n) \overset{\mathrm{d}}{=} \left(\sum_{k=1}^{n} \sum_{i=1}^{n} f(\xi_i, \dots, \xi_{i+k-1 \bmod n}) \;\Big|\; S_n = n-1 \right), \qquad (*)$$

where $j \bmod n$ denotes the index in $\{1, \dots, n\}$ that is congruent to j modulo n.

The proofs are based on this representation.

This eliminates the combinatorics, and we are left with pure probability theory!

Expectations

We calculate the expectation $\mathbb{E}\, F(\mathcal{T}_n)$ using (*), which converts this into a problem on expectations of functionals of a sequence of i.i.d. variables conditioned on their sum. (Results of this type have been studied before under various conditions.) By (*) and symmetry,

$$\mathbb{E}\, F(\mathcal{T}_n) = n \sum_{k=1}^{n} \mathbb{E}\big(f(\xi_1, \dots, \xi_k) \mid S_n = n-1 \big).$$

Let $f_k(T) := f(T) \cdot \mathbf{1}\{|T| = k\}$, and F_k the corresponding sum.

Lemma

If $1 \leq k \leq n$, then

$$\mathbb{E}\, F_k(\mathcal{T}_n) = n \frac{\mathbb{P}(S_{n-k} = n-k)}{\mathbb{P}(S_n = n-1)}\, \mathbb{E}\, f_k(\mathcal{T}).$$

The following estimates are shown using the local limit theorem and the methods used to prove it.

Lemma

Uniformly for all k with $1 \leq k \leq n/2$, as $n \to \infty$,

$$\frac{\mathbb{P}(S_{n-k} = n-k)}{\mathbb{P}(S_n = n-1)} = 1 + O\left(\frac{k}{n}\right) + o\big(n^{-1/2}\big).$$

If $n/2 < k \leq n$, then

$$\frac{\mathbb{P}(S_{n-k} = n-k)}{\mathbb{P}(S_n = n-1)} = O\left(\frac{n^{1/2}}{(n-k+1)^{1/2}}\right).$$

Variances and covariances

The arguments for variances and covariances are simila but more complicated. (More care is required since there typically is important cancellation between different terms.)
We also show a uniform bound valid for all n.

Theorem

For any functional $f : \mathfrak{T} \to \mathbb{R}$,

$$\operatorname{Var}\bigl(F(\mathcal{T}_n)\bigr)^{1/2} \le C_1 n^{1/2} \Bigl(\sup_k \sqrt{\mathbb{E}\, f(\mathcal{T}_k)^2} + \sum_{k=1}^{\infty} \frac{\sqrt{\mathbb{E}\, f(\mathcal{T}_k)^2}}{k} \Bigr),$$

with C_1 *independent of* f.

This bound is used in truncation arguments.

Asymptotic normality

We first consider functionals f with finite support. We use the representation (*), where now it suffices to sum over $k \le m$ for some $m < \infty$. We define

$$g(x_1, \dots, x_m) := \sum_{k=1}^{m} f(x_1, \dots, x_k) = \sum_{k=1}^{m} f_k(x_1, \dots, x_k).$$

Then (*) can be written (assuming $n \ge m$)

$$F(\mathcal{T}_n) \overset{\mathrm{d}}{=} \left(\sum_{i=1}^{n} g(\xi_i, \dots, \xi_{i+m-1 \bmod n}) \;\middle|\; S_n = n-1 \right).$$

Asymptotic normality now follows by a method by Le Cam (1958) and Holst (1981).

For general f we use truncations.

THE REST IS SILENCE