Distributional Results for the k-Robinson-Foulds Distance of Random Cayley Trees

(joint with Cheng-Kai Yeh and Mike Steel)

Michael Fuchs

Department of Mathematical Sciences
Chengchi University
Taipei, Taiwan

NATIONAL CHENGCHI UNIVERSITY

August 29th, 2025

Robinson-Foulds Distance of Phylogenetic Trees

T ... a phylogenetic tree.

Robinson-Foulds Distance of Phylogenetic Trees

T ... a phylogenetic tree.

Every $e \in E(T)$ gives a *split* $A|B$.

Robinson-Foulds Distance of Phylogenetic Trees

T ... a phylogenetic tree.

Every $e \in E(T)$ gives a *split* $A|B$.

Definition (RF distance)

The Robinson-Foulds distance (RF distance) $d_{RF}(T_1, T_2)$ *of* T_1 *and* T_2 *is the number of splits which only occurs either in* T_1 *or in* T_2 *but not in both.*

Robinson-Foulds Distance of Phylogenetic Trees

T ... a phylogenetic tree.

Every $e \in E(T)$ gives a *split* $A|B$.

Definition (RF distance)

The Robinson-Foulds distance (RF distance) $d_{RF}(T_1, T_2)$ of T_1 and T_2 is the number of splits which only occurs either in T_1 or in T_2 but not in both.

Question: What is the limit distribution of $d_{RF}(\mathscr{T}_1, \mathscr{T}_2)$ for random uniform trees of size n?

Robinson-Foulds Distance of Phylogenetic Trees

T ... a phylogenetic tree.

Every $e \in E(T)$ gives a *split* $A|B$.

Definition (RF distance)

The Robinson-Foulds distance (RF distance) $d_{RF}(T_1, T_2)$ of T_1 and T_2 is the number of splits which only occurs either in T_1 or in T_2 but not in both.

Question: What is the limit distribution of $d_{RF}(\mathscr{T}_1, \mathscr{T}_2)$ for random uniform trees of size n?

Theorem (Penny & Steel; 1993)

As $n \to \infty$,

$$n - 3 - \frac{d_{RF}(\mathscr{T}_1, \mathscr{T}_2)}{2} \xrightarrow{d} \text{Poisson}(1/8).$$

Cayley Trees

Cayley Trees

Definition

A (unrooted or rooted) Cayley tree is a tree with n vertices which are labeled by $\{1,\dots,n\}$.

Cayley Trees

Definition

A (unrooted or rooted) Cayley tree is a tree with n vertices which are labeled by $\{1,\ldots,n\}$.

Theorem (Cayley; 1889)

$$\#\ \text{of unrooted Cayley trees} = n^{n-2};$$

$$\#\ \text{of rooted Cayley trees} = n^{n-1}.$$

Cayley Trees

Definition

A (unrooted or rooted) Cayley tree is a tree with n vertices which are labeled by $\{1,\ldots,n\}$.

Theorem (Cayley; 1889)

$$\#\text{ of unrooted Cayley trees} = n^{n-2};$$
$$\#\text{ of rooted Cayley trees} = n^{n-1}.$$

The RF-distance was recently defined for (unrooted and rooted) Cayley's trees in:

E. Khayatian, G. Valiente, L. Zhang (2024). The k-Robinson-Foulds measure for labeled trees, Journal of Comput. Biol., 31:4, 328–344.

k-RF Distance of Cayley Trees (i)

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \le k \le n-2$.

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \leq k \leq n-2$.
- For $e = \{u, v\}$, $T - e$ has components C_u and C_v.

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \le k \le n-2$.
- For $e = \{u, v\}$, $T - e$ has components C_u and C_v.
- Define:

$$N_e(u,k) := \{w \in C_u \ : \ d(w,u) \le k\};$$
$$N_e(v,k) := \{w \in C_v \ : \ d(w,v) \le k\}.$$

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \le k \le n-2$.
- For $e = \{u, v\}$, $T - e$ has components C_u and C_v.
- Define:

$$N_e(u,k) := \{w \in C_u \;:\; d(w,u) \le k\};$$
$$N_e(v,k) := \{w \in C_v \;:\; d(w,v) \le k\}.$$

- $\{N_e(u,k), N_e(v,k)\}$ is called *k-local split*.

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \leq k \leq n-2$.
- For $e = \{u, v\}$, $T - e$ has components C_u and C_v.
- Define:

$$N_e(u,k) := \{w \in C_u \ : \ d(w,u) \leq k\};$$
$$N_e(v,k) := \{w \in C_v \ : \ d(w,v) \leq k\}.$$

- $\{N_e(u,k), N_e(v,k)\}$ is called *k-local split*.
- Let $L_k(T)$ be the set of all k-local splits.

k-RF Distance of Cayley Trees (i)

- Let T be a Cayley tree and $0 \le k \le n-2$.
- For $e = \{u, v\}$, $T - e$ has components C_u and C_v.
- Define:

$$N_e(u,k) := \{w \in C_u \ : \ d(w,u) \le k\};$$
$$N_e(v,k) := \{w \in C_v \ : \ d(w,v) \le k\}.$$

- $\{N_e(u,k), N_e(v,k)\}$ is called *k-local split*.
- Let $L_k(T)$ be the set of all k-local splits.
- For Cayley trees T_1 and T_2:

$$d_{k-RF}(T_1, T_2) := |L_k(T_1) \Delta L_k(T_2)|.$$

k-RF Distance of Cayley Trees (ii)

$S_k(T_1, T_2)$ ... number of k-local splits shared by T_1 and T_2.

k-RF Distance of Cayley Trees (ii)

$S_k(T_1, T_2)$ ... number of k-local splits shared by T_1 and T_2.

Then,

$$d_{k-RF}(T_1, T_2) = 2(n-1) - 2S_k(T_1, T_2).$$

k-RF Distance of Cayley Trees (ii)

$S_k(T_1,T_2)$ ... number of k-local splits shared by T_1 and T_2.

Then,

$$d_{k-RF}(T_1,T_2) = 2(n-1) - 2S_k(T_1,T_2).$$

Example:

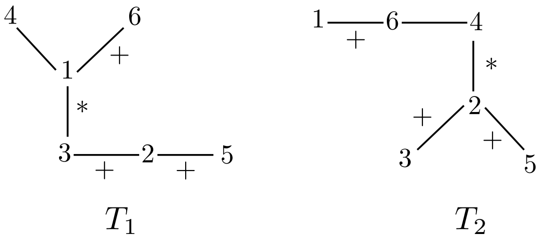
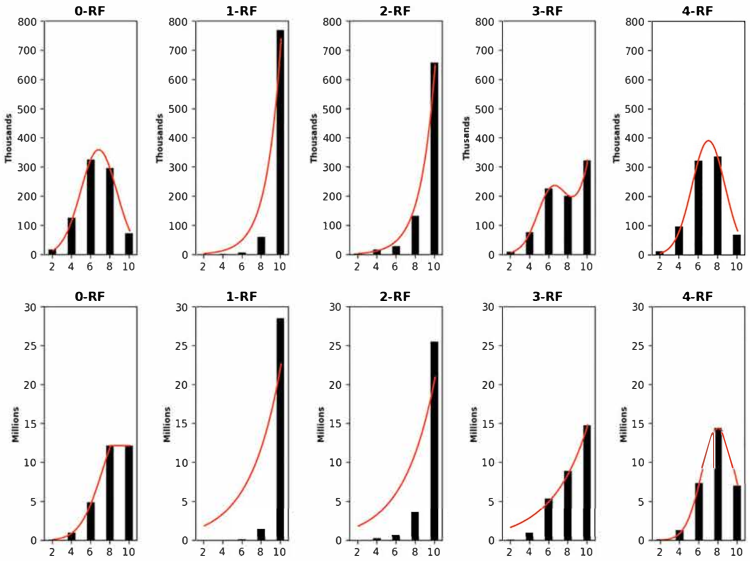


Figure: Shared k-local splits for $k=0$ (+) and $k=4$ (*).

Histogram for $n = 6$ (from Khayatian & Valiente & Zhang)



Main Result

Khayatian et al. conjectured that the limit distribution at the two boundary cases, i.e., $k = 0$ and $k = n - 2$, should be Poisson.

Main Result

Khayatian et al. conjectured that the limit distribution at the two boundary cases, i.e., $k = 0$ and $k = n-2$, should be Poisson.

Theorem (F. & Steel; 2025)

Main Result

Khayatian et al. conjectured that the limit distribution at the two boundary cases, i.e., $k=0$ and $k=n-2$, should be Poisson.

Theorem (F. & Steel; 2025)

(i) *For* $k=0$,

$$n-1-\frac{d_{0-RF}(\mathcal{T}_1,\mathcal{T}_2)}{2}\xrightarrow{d}\text{Poisson}(2).$$

Main Result

Khayatian et al. conjectured that the limit distribution at the two boundary cases, i.e., $k = 0$ and $k = n-2$, should be Poisson.

Theorem (F. & Steel; 2025)

(i) *For* $k = 0$,

$$n - 1 - \frac{d_{0-RF}(\mathcal{T}_1, \mathcal{T}_2)}{2} \xrightarrow{d} \text{Poisson}(2).$$

(ii) *For* $k = n-2$,

$$\frac{d_{(n-2)-RF}(\mathcal{T}_1, \mathcal{T}_2) - 2n(1-e^{-2})}{2\sqrt{(e^{-2} - 3e^{-4})n}} \xrightarrow{d} N(0,1).$$

Main Result

Khayatian et al. conjectured that the limit distribution at the two boundary cases, i.e., $k=0$ and $k=n-2$, should be Poisson.

Theorem (F. & Steel; 2025)

(i) *For* $k=0$,

$$n-1-\frac{d_{0-RF}(\mathscr{T}_1,\mathscr{T}_2)}{2}\xrightarrow{d}\mathrm{Poisson}(2).$$

(ii) *For* $k=n-2$,

$$\frac{d_{(n-2)-RF}(\mathscr{T}_1,\mathscr{T}_2)-2n(1-e^{-2})}{2\sqrt{(e^{-2}-3e^{-4})n}}\xrightarrow{d}N(0,1).$$

For the proof, we can equivalently work with $S_k(\mathscr{T}_1,\mathscr{T}_2)$.

$k = n - 2$ (i)

$k = n - 2$ (i)

S_n … number of splits shared by two random trees T_1 and T_2 of size n.
S'_n … number of shared trivial splits (i.e., common leaves).

$k = n-2$ (i)

S_n ... number of splits shared by two random trees T_1 and T_2 of size n.
S'_n ... number of shared trivial splits (i.e., common leaves).

Lemma

As $n \to \infty$,

$$\frac{S_n - S'_n}{\sqrt{n}} \xrightarrow{p} 0.$$

$k = n - 2$ (i)

S_n … number of splits shared by two random trees T_1 and T_2 of size n.
S'_n … number of shared trivial splits (i.e., common leaves).

Lemma

As $n \to \infty$,

$$\frac{S_n - S'_n}{\sqrt{n}} \xrightarrow{p} 0.$$

Thus the result follows from:

Proposition

As $n \to \infty$,

$$\frac{S'_n - ne^{-2}}{\sqrt{(e^{-2} - 3e^{-4})n}} \xrightarrow{d} N(0,1).$$

$k = n-2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n-1)\cdots(S'_n-k+1))$$

$k = n - 2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n - 1)\cdots(S'_n - k + 1)) = k!\binom{n}{k}$$

$k = n - 2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n - 1)\cdots(S'_n - k + 1)) = k!\binom{n}{k}\left(\frac{(n-k)^{n-k-2}(n-k)^k}{n^{n-2}}\right)^2$$

$k = n - 2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n-1)\cdots(S'_n-k+1)) = k!\binom{n}{k}\left(\frac{(n-k)^{n-k-2}(n-k)^k}{n^{n-2}}\right)^2$$
$$= k!\binom{n}{k}\left(1-\frac{k}{n}\right)^{2n-4}.$$

$k = n - 2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n - 1)\cdots(S'_n - k + 1)) = k!\binom{n}{k}\left(\frac{(n-k)^{n-k-2}(n-k)^k}{n^{n-2}}\right)^2 = k!\binom{n}{k}\left(1 - \frac{k}{n}\right)^{2n-4}.$$

Corollary

(i) $\mathbb{E}(S'_n) \sim ne^{-2}$ *and* $\mathrm{Var}(S'_n) \sim (e^{-2} - 3e^{-4})n$.

$k = n-2$ (ii)

Note:

$$\mathbb{E}(S'_n(S'_n-1)\cdots(S'_n-k+1)) = k!\binom{n}{k}\left(\frac{(n-k)^{n-k-2}(n-k)^k}{n^{n-2}}\right)^2$$
$$= k!\binom{n}{k}\left(1-\frac{k}{n}\right)^{2n-4}.$$

Corollary

(i) $\mathbb{E}(S'_n) \sim ne^{-2}$ *and* $\mathrm{Var}(S'_n) \sim (e^{-2}-3e^{-4})n$.

(ii) *For* $m \geq 1$,

$$\mathbb{E}\left(\left(\frac{S'_n - ne^{-2}}{\sqrt{(e^{-2}-3e^{-4})n}}\right)^m\right) \sim \begin{cases} m!/(2^{m/2}(m/2)!), & \textit{if } m \textit{ is even;} \\ 0, & \textit{if } m \textit{ is odd.} \end{cases}$$

$k = n - 2$ (iii)

One can also directly work with factorial moments.

$k = n - 2$ (iii)

One can also directly work with factorial moments.

Theorem (Gao & Wormald; 2004)

Let $s_n > -\mu_n^{-1}$ *and*

$$\sigma_n = \sqrt{\mu_n + \mu_n^2 s_n},$$

where $0 < \mu_n \to \infty$ *and* $\mu_n = o(\sigma_n^3)$. *Let* X_n *be a sequence of RVs with*

$$\mathbb{E}(X_n(X_n - 1)\cdots(X_n - k + 1)) \sim \mu_n^k e^{k^2 s_n/2}$$

uniformly for $c\mu_n/\sigma_n \le k \le c'\mu_n/\sigma_n$, *where* $c' > c > 0$.

Then, as $n \to \infty$,

$$\frac{X_n - \mu_n}{\sigma_n} \xrightarrow{d} N(0,1).$$

$k = n - 2$ (iv)

Other methods of proof.

$k = n - 2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

$k = n - 2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

(ii) S'_n ... number of empty boxes when throwing $2n-4$ balls uniformly at random into n urns.

$k = n - 2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

(ii) S'_n ... number of empty boxes when throwing $2n-4$ balls uniformly at random into n urns.

$\longrightarrow$ Use results for urn models.

$k = n - 2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

(ii) S'_n ... number of empty boxes when throwing $2n-4$ balls uniformly at random into n urns.

$\longrightarrow$ Use results for urn models.

(iii) Let $P_n(z)$ be the probability-generating function of S'_n.

$k = n-2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

(ii) S_n' … number of empty boxes when throwing $2n-4$ balls uniformly at random into n urns.

$\longrightarrow$ Use results for urn models.

(iii) Let $P_n(z)$ be the probability-generating function of S_n'. Then,

$$P_n(y) = \frac{(2n-4)!}{n^{2n-4}}[z^{2n-4}](e^z + y - 1)^n.$$

$k = n - 2$ (iv)

Other methods of proof.

(i) Use the (known) limit distribution result for the number of leaves a random Cayley tree.

(ii) S'_n ... number of empty boxes when throwing $2n-4$ balls uniformly at random into n urns.

$\longrightarrow$ Use results for urn models.

(iii) Let $P_n(z)$ be the probability-generating function of S'_n. Then,

$$P_n(y) = \frac{(2n-4)!}{n^{2n-4}}[z^{2n-4}](e^z + y - 1)^n.$$

Now, use saddle point method.

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2} \qquad .$$

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\qquad\qquad}{n^{n-2}}\right)^2 .$$

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 .$$

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 = \binom{n}{2}\left(\frac{2}{n}\right)^2 .$$

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 = \binom{n}{2}\left(\frac{2}{n}\right)^2 = \frac{2(n-1)}{n}.$$

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 = \binom{n}{2}\left(\frac{2}{n}\right)^2 = \frac{2(n-1)}{n}.$$

Similarly, one can compute the second factorial moments by considering the number of trees which contain two different fixed edges.

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 = \binom{n}{2}\left(\frac{2}{n}\right)^2 = \frac{2(n-1)}{n}.$$

Similarly, one can compute the second factorial moments by considering the number of trees which contain two different fixed edges.

Proposition

The number of Cayley trees which contain a spanning forest F consisting of m trees equals:

$$q_1 \cdots q_m n^{m-2},$$

where q_i denotes the number of vertices in the i-th tree in F.

$k = 0$ (i)

S_n ... number of common edges of two random trees T_1 and T_2 of size n.

Note:

$$\mathbb{E}(S_n) = \binom{n}{2}\left(\frac{\sum_{k=1}^{n-1}\binom{n-2}{k-1}k^{k-2}(n-k)^{n-k-2}}{n^{n-2}}\right)^2 = \binom{n}{2}\left(\frac{2}{n}\right)^2 = \frac{2(n-1)}{n}.$$

Similarly, one can compute the second factorial moments by considering the number of trees which contain two different fixed edges.

Proposition (Moon; 1970)

The number of Cayley trees which contain a spanning forest F consisting of m trees equals:

$$q_1 \cdots q_m n^{m-2},$$

where q_i denotes the number of vertices in the i-th tree in F.

$k = 0$ (ii)

Set:

$$S_n = \sum_{u,v} X_{u,v},$$

where $X_{u,v} = 1$ if $\{u, v\}$ is a common edge and 0 otherwise.

$k = 0$ (ii)

Set:

$$S_n = \sum_{u,v} X_{u,v},$$

where $X_{u,v} = 1$ if $\{u, v\}$ is a common edge and 0 otherwise.

$N_{u,v}$ ... contains $\{u, v\}$ and all other edges incident to $\{u, v\}$ in K_n.

$k = 0$ (ii)

Set:

$$S_n = \sum_{u,v} X_{u,v},$$

where $X_{u,v} = 1$ if $\{u, v\}$ is a common edge and 0 otherwise.

$N_{u,v}$ ... contains $\{u, v\}$ and all other edges incident to $\{u, v\}$ in K_n.

Observation: $X_{u,v}$ is independent of $(X_{r,s} \ : \ \{r, s\} \notin N_{u,v})$.

$k = 0$ (ii)

Set:

$$S_n = \sum_{u,v} X_{u,v},$$

where $X_{u,v} = 1$ if $\{u, v\}$ is a common edge and 0 otherwise.

$N_{u,v}$ ... contains $\{u, v\}$ and all other edges incident to $\{u, v\}$ in K_n.

Observation: $X_{u,v}$ is independent of $(X_{r,s} \ : \ \{r, s\} \notin N_{u,v})$.

Thus, one can use the "dissociated case" of the Stein-Chen bound to prove the following result:

Proposition

We have,

$$d_{TV}(S_n, \mathrm{Poisson}(2(n-1)/n)) = O(1/n).$$

$k = 0$ (iii)

For the rooted case, the situation is slightly different.

$k = 0$ (iii)

For the rooted case, the situation is slightly different.

Proposition (F. & Yeh)

The number of rooted Cayley trees which contain a spanning forest F consisting of m rooted trees equals:

$$n^{m-3}.$$

$k = 0$ (iii)

For the rooted case, the situation is slightly different.

Proposition (F. & Yeh)

The number of rooted Cayley trees which contain a spanning forest F consisting of m rooted trees equals:

$$n^{m-3}.$$

Theorem (F. & Yeh)

For $k = 0$,

$$n - 1 - \frac{\vec{d}_{0-RF}(\mathscr{T}_1, \mathscr{T}_2)}{2} \xrightarrow{d} \text{Poisson}(1).$$

$k = 0$ (iii)

For the rooted case, the situation is slightly different.

Proposition (F. & Yeh)

The number of rooted Cayley trees which contain a spanning forest F consisting of m rooted trees equals:

$$n^{m-3}.$$

Theorem (F. & Yeh)

For $k = 0$,

$$n - 1 - \frac{\vec{d}_{0-RF}(\mathscr{T}_1, \mathscr{T}_2)}{2} \xrightarrow{d} \text{Poisson}(1).$$

On the other hand, the result for $k = n - 2$ remains unchanged.

Summary

Summary

- Our results explain the previous simulation result.

Summary

- Our results explain the previous simulation result.
- For unrooted and rooted Cayley trees, the results are the same for $k = n - 2$ but slightly different for $k = 0$.

Summary

- Our results explain the previous simulation result.
- For unrooted and rooted Cayley trees, the results are the same for $k = n - 2$ but slightly different for $k = 0$.
- For $k = n - 3$, the limit law is again normal; for $k = 1$, the limit law is (degenerate) Poisson.

Summary

- Our results explain the previous simulation result.
- For unrooted and rooted Cayley trees, the results are the same for $k = n-2$ but slightly different for $k = 0$.
- For $k = n-3$, the limit law is again normal; for $k = 1$, the limit law is (degenerate) Poisson.

  Question: What happens for $2 \leq k \leq n-4$? Where does the limit law change from Poisson to normal?

Summary

- Our results explain the previous simulation result.
- For unrooted and rooted Cayley trees, the results are the same for $k = n-2$ but slightly different for $k = 0$.
- For $k = n-3$, the limit law is again normal; for $k = 1$, the limit law is (degenerate) Poisson.

  Question: What happens for $2 \leq k \leq n-4$? Where does the limit law change from Poisson to normal?
- In the recent paper

  E. Khayatian and L. Zhang. Simple k-RF Metrics for Comparison of Labeled DAGs, bioRxiv

  a central limit theorem is conjectured for $d_{k-s-\mathrm{RF}}(\mathscr{T}_1, \mathscr{T}_2)$ for $1 \leq k \leq n-1$.

Summary

- Our results explain the previous simulation result.
- For unrooted and rooted Cayley trees, the results are the same for $k = n-2$ but slightly different for $k = 0$.
- For $k = n-3$, the limit law is again normal; for $k = 1$, the limit law is (degenerate) Poisson.

  Question: What happens for $2 \le k \le n-4$? Where does the limit law change from Poisson to normal?
- In the recent paper

  E. Khayatian and L. Zhang. Simple k-RF Metrics for Comparison of Labeled DAGs, bioRxiv

  a central limit theorem is conjectured for $d_{k-s-\mathrm{RF}}(\mathscr{T}_1, \mathscr{T}_2)$ for $1 \le k \le n-1$. This conjecture can also be proved with our tools (joint with Bernhard Gittenberger, TU Wien).

Reference

M. Fuchs and M. Steel. The asymptotic distribution of the k-Robinson-Foulds dissimilarity measure on labeled trees, Journal of Comput. Biol., in press.

Reference

M. Fuchs and M. Steel. The asymptotic distribution of the k-Robinson-Foulds dissimilarity measure on labeled trees, Journal of Comput. Biol., in press.

Thanks for the attention!