Limit theorems for patterns in ranked tree-child networks

Tsan-Cheng Yu

(Joint work with Michael Fuchs and Hexuan Liu)

National Chengchi University

AofA, Taipei, June 28th, 2023

Studying properties of shape statistics for random models that are used to describe the evolutionary relationship between species is an important topic in biology.

For phylogenetic trees, which are used to model non-reticulate evolution, many such studies have been performed and the stochastic behavior of, e.g., pattern counts are known in great detail.

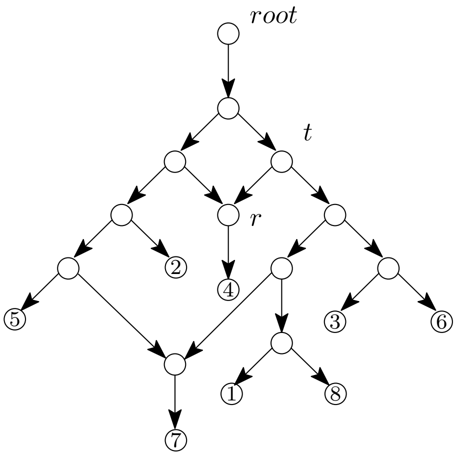
On the other hand, for phylogenetic networks, which are used to model reticulate evolution, very little is known about the number of occurrences of patterns when the networks from a given class are randomly sampled.
1999.)

Rooted binary phylogenetic networks

A rooted, binary phylogenetic network is a directed acyclic graph (DAG) without double edges such that every node falls into one of the following four categories:

- A (unique) root which has in-degree 0 and out-degree 1;
- Leaves which have in-degree 1 and out-degree 0 and which are bijectively labeled by $\{1, 2, \ldots, n\}$
- Tree nodes which are nodes of in-degree 1 and out-degree 2;
- Reticulations which are nodes of in-degree 2 and out-degree 1.

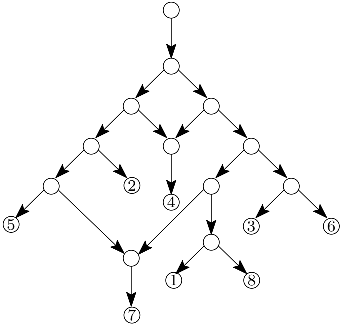
An example of rooted binary phylogenetic network



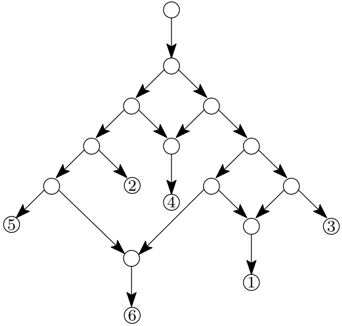
Tree-Child networks

Definition

A phylogenetic network is called *tree-child network* if every non-leaf node has at least one child which is not a reticulation.



a tree child network



a phylogenetic network but not a tree-child network

For a tree-child network, we call a tree-node a branching event and a reticulation node with its two parents a reticulation event. The vertical edges in this depiction are subsequently be called lineages.

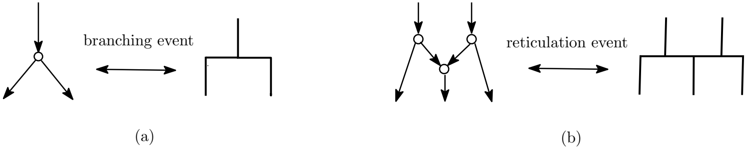
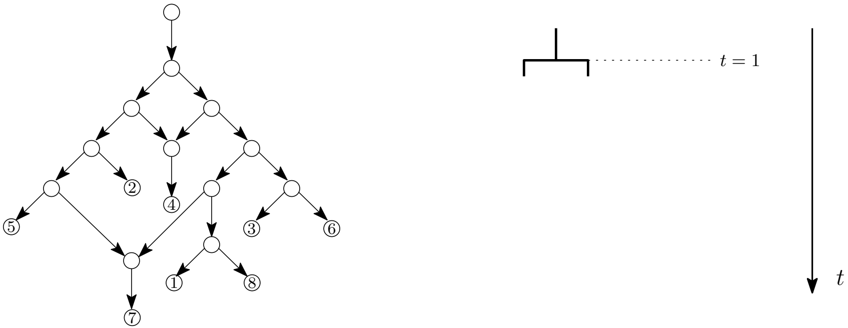


Figure: The branching and reticulation event used in the construction of ranked tree-child networks.

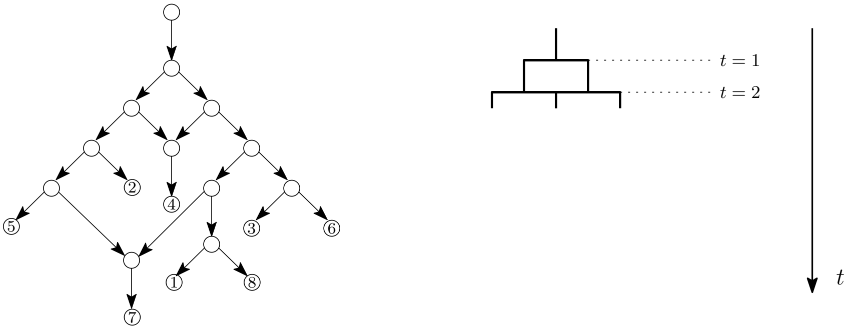
Definition

A tree-child network is called *rankable* if it has recursively evolved starting from a branching event by attaching in each step either a branching event or a reticulation event. A rankable tree-child network together with a ranking of its events is called a *ranked tree-child network (RTCN)*.



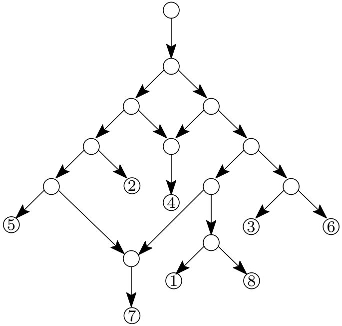
(a) (b)

Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.

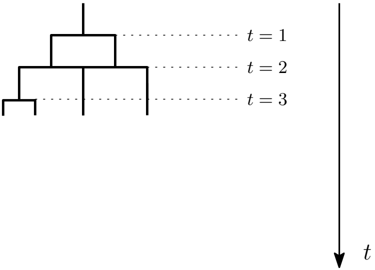


(a) (b)

Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.

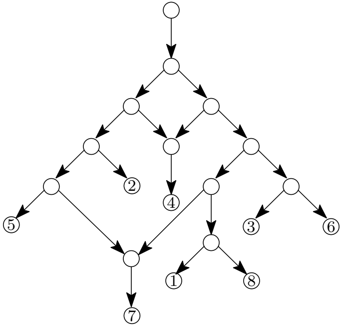


(a)



(b)

Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.



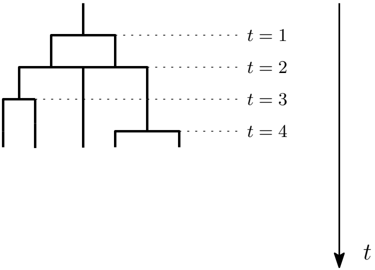
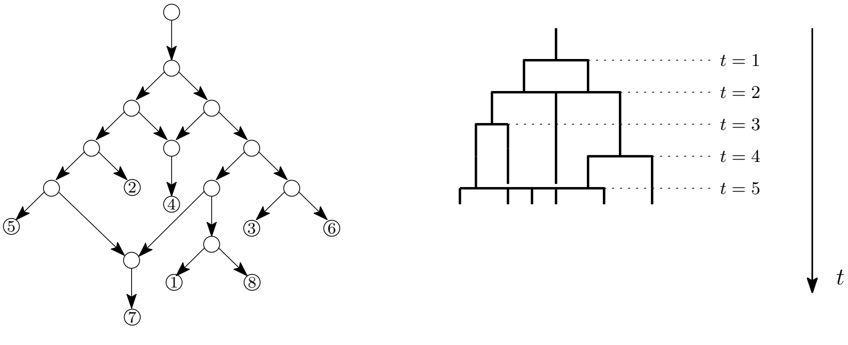
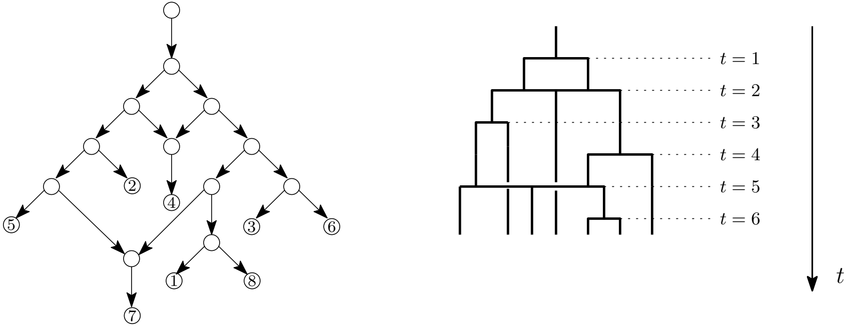


Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.



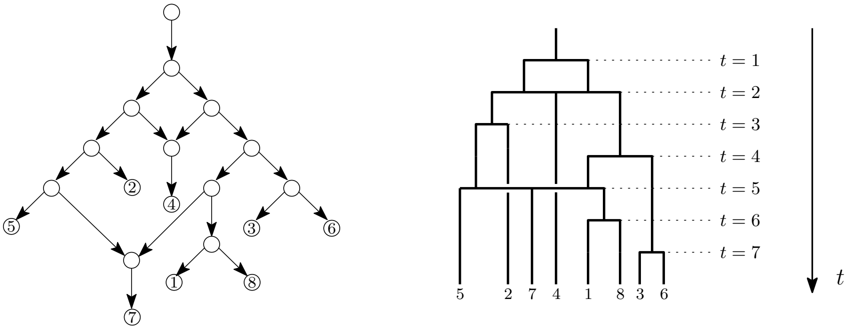
(a) (b)

Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.



(a) (b)

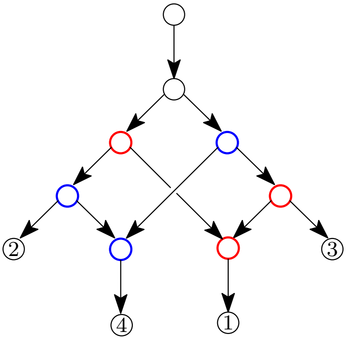
Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.



(a) (b)

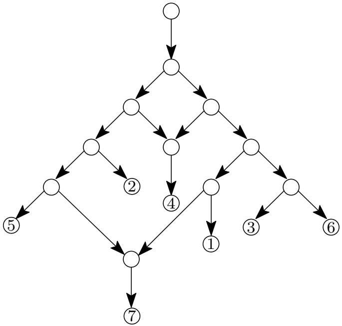
Figure: An example of rankable tree-child network and a way to have a ranked tree-child network.

A tree-child network that is not rankable.

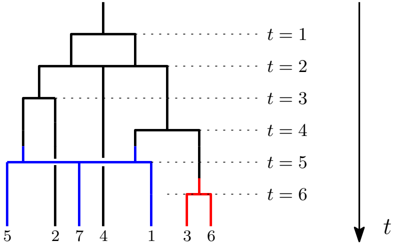


We need two definitions, when introducing results on ranked tree-child networks:

- a *cherry* is a tree node with both children leaves (or equivalently, a branching event with both outgoing lineages external);
- a trident is a reticulation event with all three outgoing lineages external.



(a)



(b)

Denote C_n (resp. T_n) the number of cherries (resp. tridents) in a uniform random ranked tree-child network with n leaves.

Theorem (Bienvenu, Lambert, and Steel in 2022)

1. *C_n weakly converges to the Poisson distribution with parameter $1/4$, i.e.,*

$$c_n \xrightarrow{d} \textit{Poisson}(1/4),$$

as $n \to \infty$.

2. $\dfrac{T_n}{n} \xrightarrow{\mathbb{P}} \dfrac{1}{7}, \ (n \to \infty),$
where $\xrightarrow{\mathbb{P}}$ denotes convergence in probability.

Theorem (Fuchs, Liu, Yu)

For the number of tridents T_n in a random ranked tree-child network with n leaves, we have

$$\frac{T_n - n/7}{\sqrt{24n/637}} \xrightarrow{d} N(0,1)$$

where $N(0,1)$ denotes the standard normal distribution.

Patterns of height 2

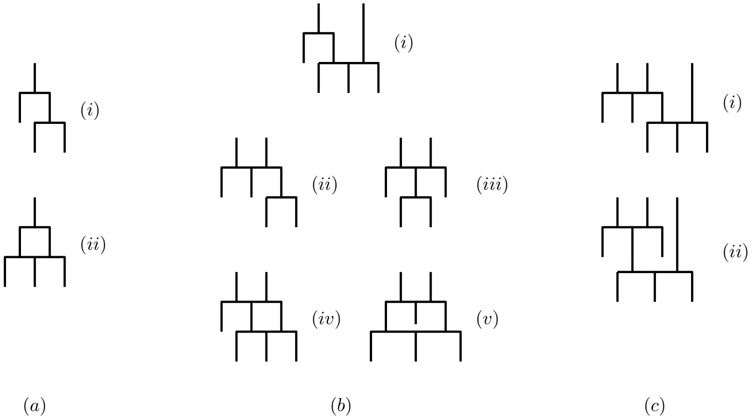


Figure: All patterns of height 2. The number of occurrences of these patterns in a ranked tree-child network with a large number of leaves is as follows: (a) These two do not occur; (b) These five occur only sporadically; (c) These two occur frequently.

Theorem (Fuchs, Liu, Yu)

Denote by X_n the number of occurrences of a (fixed) pattern of height 2 in a random ranked tree-child networks with n leaves. Then, we have the following limit law results.

(A) *For the patterns in Figure 9-(a), we have that the limit law of X_n is degenerate. More precisely,*

$$X_n \xrightarrow{L_1} 0, \qquad (n \to \infty).$$

(B) *For the patterns in Figure 9-(b), we have*

$$X_n \xrightarrow{d} \mathrm{Poisson}(\lambda), \qquad (n \to \infty),$$

where

	(b-i)	(b-ii)	(b-iii)	(b-iv)	(b-v)
λ	1/8	1/28	1/56	1/14	1/28

(C) For the patterns in Figure 9-(c), we have

$$\frac{X_n - \mu n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1), \qquad (n \to \infty),$$

where $(\mu, \sigma^2) = (4/77, 4575916/137582445)$ and $(\mu, \sigma^2) = (2/77, 2930764/137582445)$ for the patterns from Figure 9-(c-i) and Figure 9-(c-ii), respectively.

Outline of proof

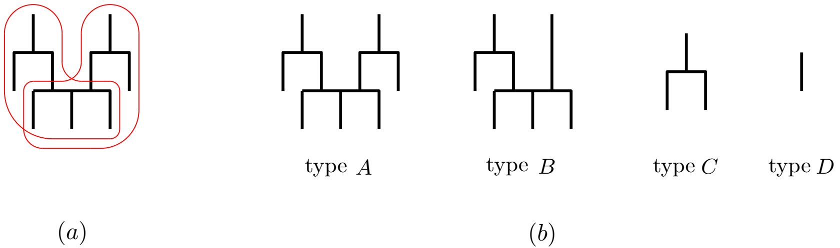
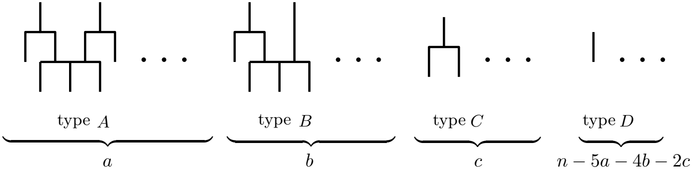


Figure: (a) The pattern of height 3 which contains two overlapping patterns from Figure 9-(b-i); (b) The types of patterns considered in the proof of the Poisson limit law for the pattern in Figure 9-(b-i); in order that every external lineage belongs exactly to one type, a pattern of type B is not allowed to be contained in a pattern of type A; also, the lineage in type D is not an external lineage in A, B or C.

Outline of proof

	type A	type B	type C	probability
A	-1	0	$+1$	$3a/n^2$
	-1	$+1$	$+1$	$2a/n^2$
B	0	-1	$+1$	$4b/n^2$
C	0	0	0	$2c/n^2$
D	0	0	$+1$	$(n-5a-4b-2c)/n^2$

Table: The change of the number of patterns of type A, type B and type C (see Figure 10-(b)) when the next event is a branching event.



Outline of proof

	type A	type B	type C	probability
A	-1	0	0	$20an^2$
A & A	-2	0	0	$9a(a-1)/n^2$
	-2	$+1$	0	$12a(a-1)/n^2$
	-2	$+2$	0	$4a(a-1)/n^2$
B	0	-1	0	$12b/n^2$
B& B	0	-2	0	$16b(b-1)/n^2$
C	0	0	-1	$2c/n^2$
C & C	$+1$	0	-2	$4c(c-1)/n^2$
D & D	0	0	0	$(n-5a-4b-2c)(n-5a-4b-2c-1)/n^2$
A & B	-1	-1	0	$24ab/n^2$
	-1	0	0	$16ab/n^2$
A & C	-1	$+1$	-1	$12ac/n^2$
	-1	$+2$	-1	$8ac/n^2$
A & D	-1	0	0	$6a(n-5a-4b-2c)/n^2$
	-1	$+1$	0	$4a(n-5a-4b-2c)/n^2$
B & C	0	0	-1	$16bc/n^2$
B & D	0	-1	0	$8b(n-5a-4b-2c)/n^2$
C & D	0	$+1$	-1	$4c(n-5a-4b-2c)/n^2$

Figure: The change of the number of patterns of type A and type B (see Figure 10-(b)) when the next event is a reticulation event.

Outline of proof

Lemma

Denote by Y_n and $\tilde{X}_n$ the number of occurrences of patterns of type A and type B, respectively, in a random ranked tree-child network with n leaves. Then, for $r, s, t \geq 0$, we have

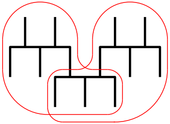
$$\mathbb{E}(Y_{n+1}^r \tilde{X}_{n+1}^s C_{n+1}^t) = \left(1 - \frac{5r + 4s + 2t}{n}\right)^2 \mathbb{E}(Y_n^r \tilde{X}_n^s C_n^t) + \frac{t}{n}\mathbb{E}(Y_n^r \tilde{X}_n^s C_n^{t-1}) + \frac{4s}{n}\mathbb{E}(Y_n^r \tilde{X}_n^{s-1} C_n^{t+1}) + \frac{R_n}{n^2}, \quad (1)$$

Outline of proof

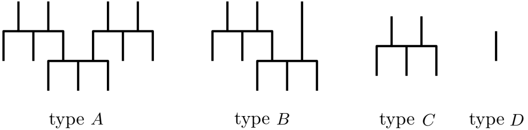
where R_n is given by

$$
\begin{aligned}
& t(2-5r-4s-2t)\mathbb{E}(Y_n^r \tilde{X}_n^s C_n^{t-1}) + 4r\mathbb{E}(Y_n^{r-1}\tilde{X}_n^s C_n^{t+2}) \\
& - 2s(1+10r+8s+4t)\mathbb{E}(Y_n^{r+1}\tilde{X}_n^{s-1}C_n^t) + 2st\mathbb{E}(Y_n^{r+1}\tilde{X}_n^{s-1}C_n^{t-1}) \\
& - 8s\mathbb{E}(Y_n^r\tilde{X}_n^{s-1}C_n^{t+2}) + 4s(2-5r-4s-2t)\mathbb{E}(Y_n^r\tilde{X}_n^{s-1}C_n^{t+1}) \\
& + 4s(s-1)\mathbb{E}(Y_n^{r+2}\tilde{X}_n^{s-2}C_n^t) + 8s(s-1)\mathbb{E}(Y_n^{r+1}\tilde{X}_n^{s-2}C_n^{t+1}).
\end{aligned}
$$

Outline of proof



(a)



(b)

Figure: (a) The pattern of height 3 which contains two overlapping patterns from Figure 9-(c-i); (b) The types of pattern considered in the proof of the normal limit law for the pattern in Figure 9-(c-i); in order that every external lineage belongs exactly to one type, a pattern of type B resp. C is not allowed to be contained in a pattern of type A resp. B and A; also, the lineage in type D is not an external lineage in A, B or C.

Outline of proof

	type A	type B	type C	probability
A	-1	0	0	$3a/n^2$
	-1	$+1$	0	$4a/n^2$
B	0	-1	0	$3b/n^2$
	0	-1	$+1$	$2b/n^2$
C	0	0	-1	$3c/n^2$
D	0	0	0	$(n-7a-5b-3c)/n^2$

Table: The change of the number of patterns of type A, B and C (see Figure 12-(b)) when the next event is a branching event.

Outline of proof

	type A	type B	type C	probability
A	-1	0	$+1$	$14a/n^2$
	-1	0	$+2$	$8a/n^2$
	-1	$+1$	0	$16a/n^2$
	-1	$+1$	$+1$	$4a/n^2$
A & A	-1	0	0	$4a(a-1)/n^2$
	-2	0	$+1$	$a(a-1)/n^2$
	-2	$+1$	0	$4a(a-1)/n^2$
	-2	$+1$	$+1$	$8a(a-1)/n^2$
	-2	$+2$	0	$16a(a-1)/n^2$
	-2	$+2$	$+1$	$16a(a-1)/n^2$
B	0	0	0	$8b/n^2$
	0	-1	$+1$	$10b/n^2$
	0	-1	$+2$	$2b/n^2$
B & B	0	-1	0	$4b(b-1)/n^2$
	0	-1	$+1$	$8b(b-1)/n^2$
	0	-2	$+1$	$b(b-1)/n^2$
	0	-2	$+2$	$4b(b-1)/n^2$
	0	-2	$+3$	$4b(b-1)/n^2$
	$+1$	-2	0	$4b(b-1)/n^2$
C	0	0	0	$6c/n^2$
C & C	0	0	-1	$c(c-1)/n^2$
	0	$+1$	-2	$4c(c-1)/n^2$
	$+1$	0	-2	$4c(c-1)/n^2$
D & D	0	0	$+1$	$(n-7a-5b-3c)(n-7a-5b-3c-1)/n^2$

Figure: The change of the number of patterns of type A, B and C (see Figure 12-(b)) when the next event is a reticulation event which is attached to one or two patterns of type X with $X \in \{A, B, C, D\}$.

Outline of proof

Lemma

Denote by X_n and Y_n the number of occurrences of the patterns from Figure 9-(c-i) and Figure 12-(a), respectively, in a random ranked tree-child network with n leaves. Moreover, set $\mu_n := \mathbb{E}(T_n), \rho_n := \mathbb{E}(X_n), \tau_n := \mathbb{E}(Y_n)$ and $\bar{T}_n := T_n - \mu_n, \bar{X}_n := X_n - \rho_n, \bar{Y}_n := Y_n - \tau_n$. Then, for all $r, s, t \geq 0$, we have

$$\mathbb{E}(\bar{Y}_{n+1}^r \bar{X}_{n+1}^s \bar{T}_{n+1}^t) = \left(1 - \frac{7r + 5s + 3t}{n}\right)^2 \mathbb{E}(\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^t) + R_n \quad (2)$$

with

$$R_n = \sum_{(s',r',t')} \mathbb{E}(\bar{Y}_n^{r'} \bar{X}_n^{s'} \bar{T}_n^{t'}) \Lambda_{r',s',t'}(n), \quad (3)$$

where the sum runs over (s', r', t') which are of a smaller lexicographic order than (s, r, t) and $\Lambda_{r',s',t'}(n)$ admits the complete asymptotic expansion:

Outline of proof

$$\Lambda_{r',s',t'}(n) \sim \sum_{\ell=0}^{\infty} \frac{\lambda_{r',s',t',\ell}}{n^\ell}, \qquad (n \to \infty). \tag{4}$$

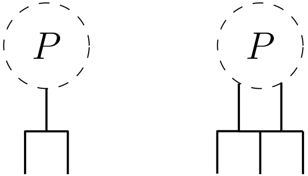
Moreover, all terms in (3) with
$(r' + s' + t')/2 - \ell \geq (r + s + t)/2 - 1$ are given by

$$\begin{aligned}
&\frac{4s}{n}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-1} \bar{T}_n^{t+1}) + \frac{8r}{7n}\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^s \bar{T}_n^{t+1}) + \binom{r}{2}\frac{80092}{540225}\mathbb{E}(\bar{Y}_n^{r-2} \bar{X}_n^s \bar{T}_n^t) \\
&\quad + \binom{s}{2}\frac{21916}{29645}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-2} \bar{T}_n^t) + \binom{t}{2}\frac{24}{49}\mathbb{E}(\bar{Y}_n^r \bar{X}_n^s \bar{T}_n^{t-2}) \\
&\quad - \frac{128}{539} st\mathbb{E}(\bar{Y}_n^r \bar{X}_n^{s-1} \bar{T}_n^{t-1}) - \frac{32}{343} rt\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^s \bar{T}_n^{t-1}) \\
&\quad + \frac{712}{3773} rs\mathbb{E}(\bar{Y}_n^{r-1} \bar{X}_n^{s-1} \bar{T}_n^t).
\end{aligned} \tag{5}$$

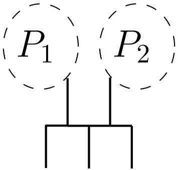
Conjecture

Let F be a fringe pattern. Denote by P resp. P_1 and P_2 the patterns which are obtained from it by removing the last event. (Here, the second case is only possible if the last event is a reticulation event and the pattern gets disconnected when this event is removed.) Then, we have the following cases.

(a) *If P is a normal pattern, then F is a Poisson pattern; in all other cases for P, the pattern F is a degenerate pattern.*



(b) If P_1 and P_2 are both normal patterns, then F is also a normal pattern; if P_1 is a normal pattern and P_2 is a Poisson pattern or vice versa, then F is a Poisson pattern; in all remaining cases for P_1 and P_2, F is a degenerate pattern.



Thank you!