Enumeration and Stochastic Properties of Tree-Child Networks

(based on joint work with Y.-S. Chang, H. Liu, M. Wallner, G.-R. Yu and L. Zhang)

Michael Fuchs

Department of Mathematical Sciences
National Chengchi University

February 8th, 2023

What is a (Binary) Phylogenetic Network?

X . . . a finite set.

What is a (Binary) Phylogenetic Network?

X ... a finite set.

Definition

A phylogenetic network is a rooted DAG with the following nodes:

What is a (Binary) Phylogenetic Network?

X ... a finite set.

Definition

A phylogenetic network is a rooted DAG with the following nodes:

(a) *root: in-degree* 0 *and out-degree* 1*;*

What is a (Binary) Phylogenetic Network?

X . . . a finite set.

Definition

A phylogenetic network *is a rooted DAG with the following nodes:*

(a) *root: in-degree* 0 *and out-degree* 1;

(b) *leaves: in-degree* 1 *and out-degree* 0; *bijectively labeled by* X;

What is a (Binary) Phylogenetic Network?

X ... a finite set.

Definition

A phylogenetic network is a rooted DAG with the following nodes:

(a) *root: in-degree 0 and out-degree 1;*

(b) *leaves: in-degree 1 and out-degree 0; bijectively labeled by X;*

(c) *all other nodes have either out-degree 2 and in-degree 1 (tree nodes) or out-degree 1 and in-degree 2 (reticulation nodes).*

What is a (Binary) Phylogenetic Network?

X ... a finite set.

Definition

A phylogenetic network is a rooted DAG with the following nodes:

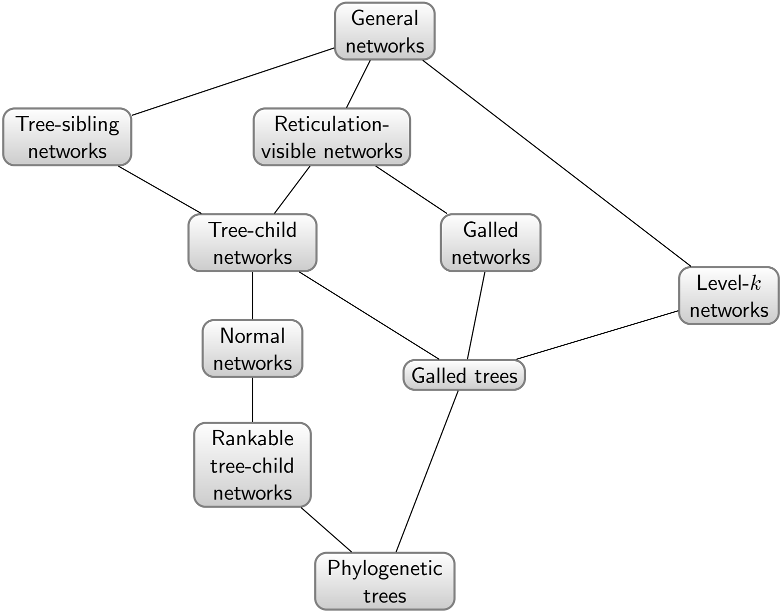
(a) *root: in-degree* 0 *and out-degree* 1*;*

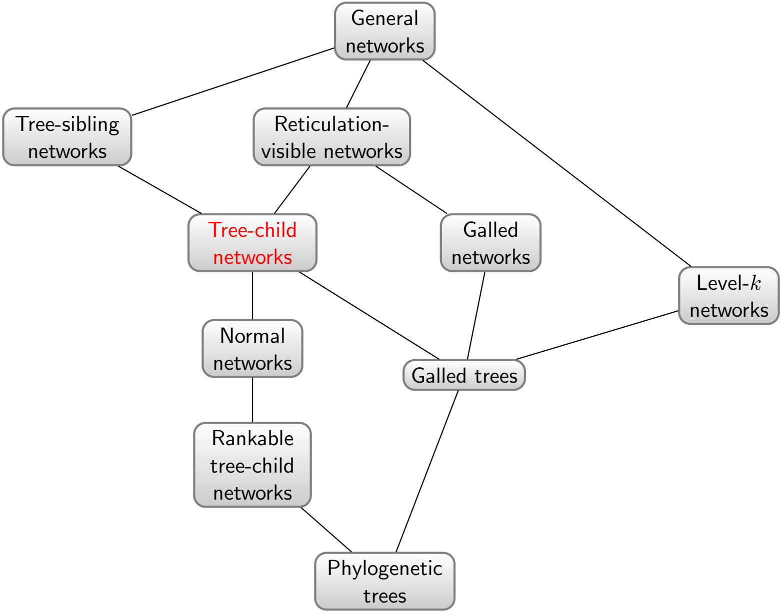
(b) *leaves: in-degree* 1 *and out-degree* 0*; bijectively labeled by* X*;*

(c) *all other nodes have either out-degree* 2 *and in-degree* 1 *(tree nodes) or out-degree* 1 *and in-degree* 2 *(reticulation nodes).*

Phylogenetic networks have become increasingly popular in recent decades.

They are used to model reticulate evolution which contains reticulation events such as lateral gene transfer or hybridization.


General networks
Tree-sibling networks
Reticulation-visible networks
Tree-child networks
Galled networks
Level-k networks
Normal networks
Galled trees
Rankable tree-child networks
Phylogenetic trees


General networks
Tree-sibling networks
Reticulation-visible networks
Tree-child networks
Galled networks
Level-k networks
Normal networks
Galled trees
Rankable tree-child networks
Phylogenetic trees

TC-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 node.

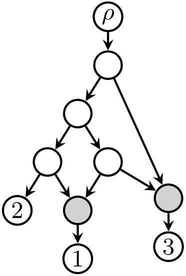
TC-Networks

Definition

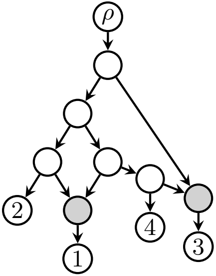
A phylogenetic network is called tree-child network if every non-leaf node has at least one child which is not a reticulation node.

Examples:

(a)



(b)



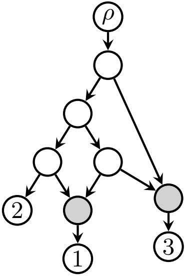
TC-Networks

Definition

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

Examples:

(a)



(b)

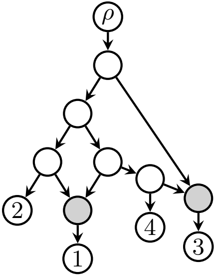


Figure: (a) is not a tc-network whereas (b) is a tc-network.

Counting TC-Networks

TC_n ... $\#$ of tc-networks with n leaves.

Counting TC-Networks

$\mathrm{TC}_n \ldots \#$ of tc-networks with n leaves.

Theorem (McDiarmid, Semple, Welsh; 2015)

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq \mathrm{TC}_n \leq (c_2 n)^{2n}.$$

Counting TC-Networks

TC_n ... # of tc-networks with n leaves.

Theorem (McDiarmid, Semple, Welsh; 2015)

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq \mathrm{TC}_n \leq (c_2 n)^{2n}.$$

Question: what is the exponential growth rate?

Counting TC-Networks

$\mathrm{TC}_n \ldots \#$ of tc-networks with n leaves.

Theorem (McDiarmid, Semple, Welsh; 2015)

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq \mathrm{TC}_n \leq (c_2 n)^{2n}.$$

Question: what is the exponential growth rate?

McDiarmid & Semple & Welsh (2015) also proved stochastic results.

Counting TC-Networks

$\mathrm{TC}_n \ldots \#$ of tc-networks with n leaves.

Theorem (McDiarmid, Semple, Welsh; 2015)

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq \mathrm{TC}_n \leq (c_2 n)^{2n}.$$

Question: what is the exponential growth rate?

McDiarmid & Semple & Welsh (2015) also proved stochastic results.

Theorem (McDiarmid, Semple, Welsh; 2015)

(a) *$\#$ of reticulation nodes $\sim n$ for almost all tc-networks;*

(b) *The number of cherries is $o(n)$ for almost all tc-networks.*

$\mathrm{TC}_{n,k}$ for small n, k (i)

$\mathrm{TC}_{n,k}$ ... # of tc-networks with n leaves and k reticulation nodes.

$\mathrm{TC}_{n,k}$ for small n, k (i)

$\mathrm{TC}_{n,k}$ ... # of tc-networks with n leaves and k reticulation nodes.

Cardona & Zhang (2020):

$k \setminus n$	2	3	4	5	6	7
1	2	21	228	2805	39330	623385
2		42	1272	30300	696600	16418430
3			2544	154500	6494400	241204950
4				309000	31534200	2068516800
5					63068400	9737380800
6						19474761600

$\mathrm{TC}_{n,k}$ for small n, k (i)

$\mathrm{TC}_{n,k}$ ... # of tc-networks with n leaves and k reticulation nodes.

Cardona & Zhang (2020):

$k \setminus n$	2	3	4	5	6	7
1	2	21	228	2805	39330	623385
2		42	1272	30300	696600	16418430
3			2544	154500	6494400	241204950
4				309000	31534200	2068516800
5					63068400	9737380800
6						19474761600

Computation becomes hard for large n with the method of Cardona & Zhang which uses component graphs whose number increases rapidly!

$\mathrm{TC}_{n,k}$ for small n, k (ii)

Pons & Batle (2021) found a recursive formula for $\mathrm{TC}_{n,k}$ based on a (still unproven) conjecture.

$\mathrm{TC}_{n,k}$ for small n, k (ii)

Pons & Batle (2021) found a recursive formula for $\mathrm{TC}_{n,k}$ based on a (still unproven) conjecture.

Chang & Liu & F. & Wallner & Yu (2023+) recently also found the following recursive formula:

$$\mathrm{TC}_{n,k} = \frac{n!}{2^{n-1-k}} w_{n-1,k},$$

where

$$\omega_{n,k} = \sum_{m\geq 1} b_{n,k,m}$$

with $b_{n,km}$ given recursively by:

$$b_{n,k,m} = \sum_{j=1}^{m} b_{n-1,k,j} + (n+m+k-2) \sum_{j=1}^{m} b_{n-1,k-1,j}.$$

A Sequence of Words

Definition (OEIS; A213863)

Denote by a_n the number of words on letters $\{\omega_1, \dots, \omega_n\}$ so that

(i) *each letter occurs exactly 3 times;*

(ii) *ω_i has either not occurred or it has occurred at least as often as ω_j with $j > i$.*

A Sequence of Words

Definition (OEIS; A213863)

Denote by a_n *the number of words on letters* $\{\omega_1, \ldots, \omega_n\}$ *so that*

(i) *each letter occurs exactly* 3 *times;*

(ii) ω_i *has either not occurred or it has occurred at least as often as* ω_j *with* $j > i$.

For example, $a_2 = 7$ because

$$aaabbb, aababb, aabbab, abaabb, ababab, baaabb, baabab.$$

A Sequence of Words

Definition (OEIS; A213863)

Denote by a_n *the number of words on letters* $\{\omega_1, \ldots, \omega_n\}$ *so that*

(i) *each letter occurs exactly* 3 *times;*

(ii) ω_i *has either not occurred or it has occurred at least as often as* ω_j *with* $j > i$.

For example, $a_2 = 7$ because

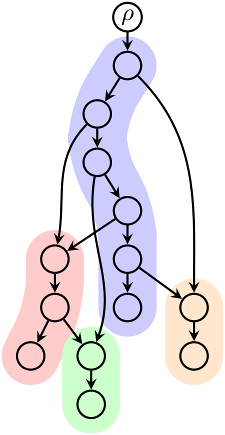
$$aaabbb, aababb, aabbab, abaabb, ababab, baaabb, baabab.$$

Proposition (F., Yu, Zhang; 2021)

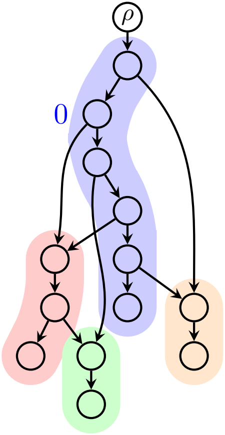
We have,

$$\frac{\mathrm{TC}_{n,n-1}}{n!} = a_{n-1}.$$

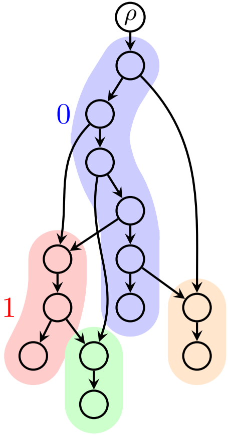
Bijective Proof



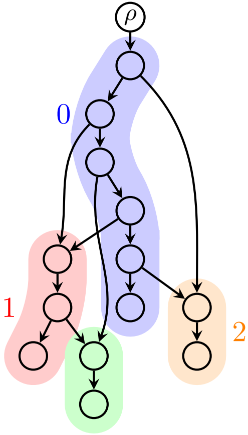
Bijective Proof



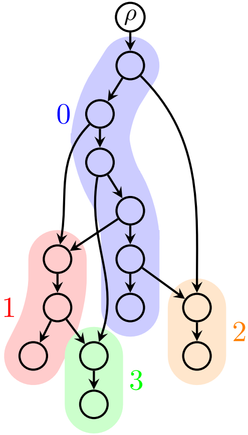
Bijective Proof



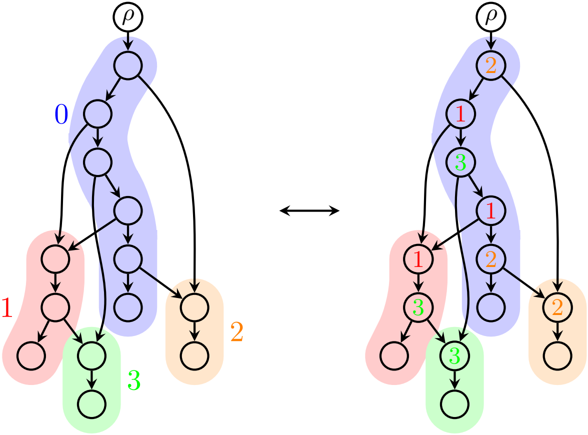
Bijective Proof



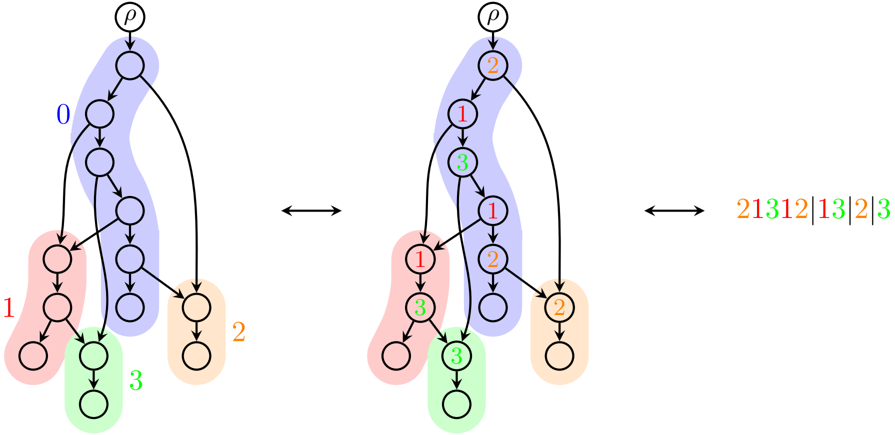
Bijective Proof



Bijective Proof



Bijective Proof



Asymptotic Counting of TC-Networks

Define

$$b_{n,m} = (2n+m-2)\sum_{j=1}^{m} b_{n-1,j}$$

Then, $a_n = \sum_{m\geq 1} b_{n,m}$.

Asymptotic Counting of TC-Networks

Define

$$b_{n,m} = (2n+m-2)\sum_{j=1}^{m} b_{n-1,j}$$

Then, $a_n = \sum_{m\geq 1} b_{n,m}$.

From this, by a recent method of Elvey Price & Fang & Wallner (2021):

Asymptotic Counting of TC-Networks

Define

$$b_{n,m} = (2n+m-2)\sum_{j=1}^{m} b_{n-1,j}$$

Then, $a_n = \sum_{m\geq 1} b_{n,m}$.

From this, by a recent method of Elvey Price & Fang & Wallner (2021):

Theorem (F., Yu, Zhang; 2021)

We have,

$$\mathrm{TC}_n = \Theta(\mathrm{TC}_{n,n-1}) = \Theta\left(n^{-2/3}e^{a_1(3n)^{1/3}}\left(\frac{12}{e^2}\right)^n n^{2n}\right),$$

where a_1 is the largest root of the Airy function of first order.

Asymptotic Counting of TC-Networks

Define

$$b_{n,m} = (2n + m - 2) \sum_{j=1}^{m} b_{n-1,j}$$

Then, $a_n = \sum_{m \geq 1} b_{n,m}$.

From this, by a recent method of Elvey Price & Fang & Wallner (2021):

Theorem (F., Yu, Zhang; 2021)

We have,

$$\mathrm{TC}_n = \Theta(\mathrm{TC}_{n,n-1}) = \Theta\left(n^{-2/3} e^{a_1 (3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\right),$$

where a_1 is the largest root of the Airy function of first order.

So, the base of the exponential growth rate is $12/e^2$!

Stochastic Results for TC-Networks

Chang & F. & Liu & Wallner & Yu (2023+) proved that

$$\mathrm{TC}_{n,n-1-k} \approx \frac{1}{2^k k!}\mathrm{TC}_{n,n-1}$$

for k close to n.

Stochastic Results for TC-Networks

Chang & F. & Liu & Wallner & Yu (2023+) proved that

$$\mathrm{TC}_{n,n-1-k} \approx \frac{1}{2^k k!}\mathrm{TC}_{n,n-1}$$

for k close to n.

Theorem (Chang, F., Liu, Wallner, Yu; 2023+)

(a) *We have,*

$$n - 1 - \#\ \textit{of reticulation nodes} \xrightarrow{d} \textit{Poisson}(1/2).$$

Stochastic Results for TC-Networks

Chang & F. & Liu & Wallner & Yu (2023+) proved that

$$\mathrm{TC}_{n,n-1-k} \approx \frac{1}{2^k k!}\mathrm{TC}_{n,n-1}$$

for k close to n.

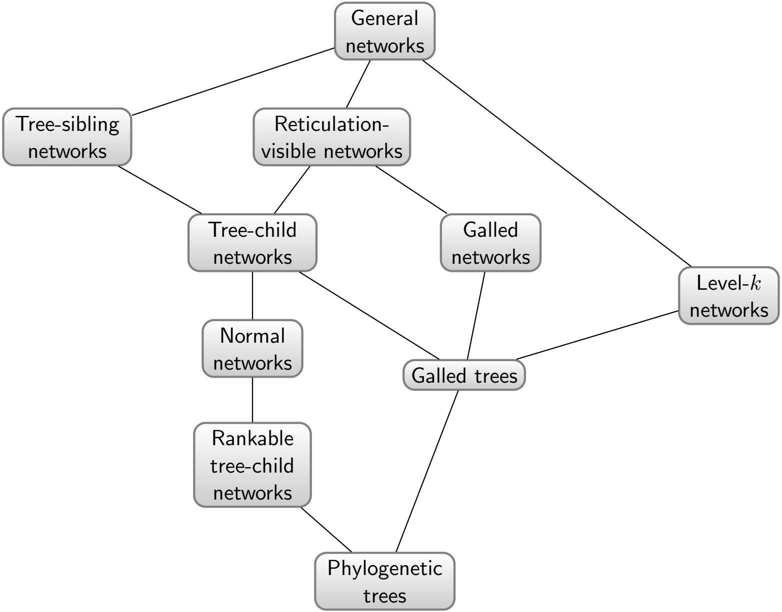
Theorem (Chang, F., Liu, Wallner, Yu; 2023+)

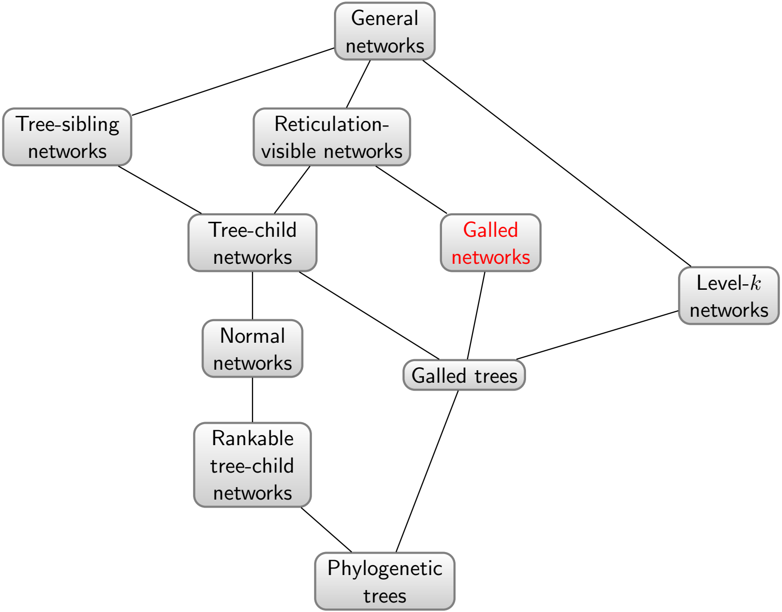
(a) *We have,*

$$n-1-\#\ \textit{of reticulation nodes} \xrightarrow{d} \textit{Poisson}(1/2).$$

(b) *We have,*

$$\mathbb{E}(\#\ \textit{of cherries}) = \mathcal{O}(1).$$


General networks
Tree-sibling networks
Reticulation-visible networks
Tree-child networks
Galled networks
Level-k networks
Normal networks
Galled trees
Rankable tree-child networks
Phylogenetic trees


General networks
Tree-sibling networks
Reticulation-visible networks
Tree-child networks
Galled networks
Level-k networks
Normal networks
Galled trees
Rankable tree-child networks
Phylogenetic trees

Results for Galled Networks

GN_n . . . # of galled networks with n leaves.

Results for Galled Networks

GN_n ... # of galled networks with n leaves.

Theorem (F., Yu, Zhang; 2022)

We have,

$$\mathrm{GN}_n \sim \frac{\sqrt{2e\sqrt[4]{e}}}{4} n^{-1} \left(\frac{8}{e^2}\right)^n n^{2n}.$$

Results for Galled Networks

GN_n … # of galled networks with n leaves.

Theorem (F., Yu, Zhang; 2022)

We have,

$$\mathrm{GN}_n \sim \frac{\sqrt{2e\sqrt[4]{e}}}{4} n^{-1} \left(\frac{8}{e^2}\right)^n n^{2n}.$$

F. & Yu & Zhang (2022) also found the limiting distribution of the number X_n of reticulation nodes. In particular,

$$\mathbb{E}(X_n) = n - \frac{3}{8} + o(1) \qquad \text{and} \qquad \mathrm{Var}(X_n) = \frac{3}{4} + o(1).$$

Open Questions

Open Questions

- How to prove the conjecture of Pons & Batle (2021)?

$$(n-k)\mathrm{TC}_{n,k} = (n+1-k)(n-k)\mathrm{TC}_{n,k-1} + n(2n+k-3)\mathrm{TC}_{n-1,k}$$

Open Questions

- How to prove the conjecture of Pons & Batle (2021)?

$$(n-k)\mathrm{TC}_{n,k} = (n+1-k)(n-k)\mathrm{TC}_{n,k-1} + n(2n+k-3)\mathrm{TC}_{n-1,k}$$

- Does there exist a constant γ such that

$$\mathrm{TC}_n \sim \gamma n^{-2/3} e^{a_1 (3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}?$$

Open Questions

- How to prove the conjecture of Pons & Batle (2021)?

$$(n-k)\mathrm{TC}_{n,k} = (n+1-k)(n-k)\mathrm{TC}_{n,k-1} + n(2n+k-3)\mathrm{TC}_{n-1,k}$$

- Does there exist a constant γ such that

$$\mathrm{TC}_n \sim \gamma n^{-2/3} e^{a_1(3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}?$$

- Further stochastic results for random tc-networks? E.g., for some results on the Sackin index see:

L. Zhang (2022). The Sackin index of simplex networks, In: Jin, L., Durand, D. (eds) Comparative Genomics. RECOMB-CG 2022, Lecture Notes in Computer Science, 13234, Springer, Cham.

Open Questions

- How to prove the conjecture of Pons & Batle (2021)?

$$(n-k)\mathrm{TC}_{n,k} = (n+1-k)(n-k)\mathrm{TC}_{n,k-1} + n(2n+k-3)\mathrm{TC}_{n-1,k}$$

- Does there exist a constant γ such that

$$\mathrm{TC}_n \sim \gamma n^{-2/3} e^{a_1 (3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}?$$

- Further stochastic results for random tc-networks? E.g., for some results on the Sackin index see:

  L. Zhang (2022). The Sackin index of simplex networks, In: Jin, L., Durand, D. (eds) Comparative Genomics. RECOMB-CG 2022, Lecture Notes in Computer Science, 13234, Springer, Cham.

- Enumeration and stochastic results for reticulation-visible networks?

References

1. G. Cardona and L. Zhang (2020). Counting and enumerating tree-child networks and their subclasses, *J. Comput. System Sci.*, **114**, 84–104.
2. Y.-S. Chang, M. Fuchs, H. Liu, M. Wallner, G.-R. Yu. Enumerative and distributional results for d-combining tree-child networks, submitted.
3. M. Fuchs, G.-R. Yu, L. Zhang (2021). On the asymptotic growth of the number of tree-child networks, *Eur. J. Combin.*, **93**, 103278.
4. M. Fuchs, G.-R. Yu, L. Zhang (2022). Asymptotic enumeration and distributional properties of galled networks, *J. Comb. Theory Ser. A.*, **189**, 105599.
5. C. McDiarmid, C. Semple, D. Welsh (2015). Counting phylogenetic networks, *Ann. Comb.*, **19:1**, 205–224.
6. M. Pons and J. Batle (2021). Combinatorial characterization of a certain class of words and a conjectured connection with general subclasses of phylogenetic tree-child networks, *Sci. Rep.*, 11, Article number: 21875.