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?

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

- Root: in-degree 0 and out-degree 1
- Leaves: in-degree 1 and out-degree 0; bijectively labeled by $X$
- 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.
What is a (Binary) Phylogenetic Network?

$X \ldots$ 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?

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

(a) \textit{root}: in-degree 0 and out-degree 1;

(b) \textit{leaves}: in-degree 1 and out-degree 0; bijectively labeled by \(X\);
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?

$X \ldots$ 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.
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.
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) \[\rho \ 2 \ 1 \ 3\]

(b) \[\rho \ 4 \ 2 \ 1 \ 3\]

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

\[ TC_n \ldots \# \text{ of tc-networks with } n \text{ leaves.} \]
Counting TC-Networks

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

**Theorem (McDiarmid, Semple, Welsh; 2015)**

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq TC_n \leq (c_2 n)^{2n}.$$
Counting TC-Networks

$TC_n \ldots \# \text{ of tc-networks with } n \text{ leaves.}$

**Theorem (McDiarmid, Semple, Welsh; 2015)**

*For constants* $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq TC_n \leq (c_2 n)^{2n}.$$ 

**Question:** what is the exponential growth rate?
Counting TC-Networks

\[ \text{TC}_n \ldots \# \text{ of tc-networks with } n \text{ leaves.} \]

**Theorem (McDiarmid, Semple, Welsh; 2015)**

For constants \(0 < c_1 < c_2\),

\[
(c_1 n)^{2n} \leq \text{TC}_n \leq (c_2 n)^{2n}.
\]

**Question:** what is the exponential growth rate?

McDiarmid & Semple & Welsh (2015) also proved stochastic results.
Counting TC-Networks

$TC_n \ldots \# \text{ of tc-networks with } n \text{ leaves.}$

**Theorem (McDiarmid, Semple, Welsh; 2015)**

For constants $0 < c_1 < c_2$,

$$(c_1 n)^{2n} \leq TC_n \leq (c_2 n)^{2n}.$$  

**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.
$\text{TC}_{n,k}$ for small $n, k$ (i)

$\text{TC}_{n,k}$ ... # of tc-networks with $n$ leaves and $k$ reticulation nodes.
**$\text{TC}_{n,k}$ for small $n, k$ (i)**

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

Cardona & Zhang (2020):

<table>
<thead>
<tr>
<th>$k \setminus n$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>21</td>
<td>228</td>
<td>2805</td>
<td>39330</td>
<td>623385</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1272</td>
<td>30300</td>
<td>696600</td>
<td>16418430</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2544</td>
<td>154500</td>
<td>6494400</td>
<td>241204950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>309000</td>
<td>31534200</td>
<td>2068516800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>63068400</td>
<td>9737380800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>19474761600</td>
<td></td>
</tr>
</tbody>
</table>
$\text{TC}_{n,k}$ for small $n$, $k$ (i)

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

Cardona & Zhang (2020):

<table>
<thead>
<tr>
<th>$k \backslash n$</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>21</td>
<td>228</td>
<td>2805</td>
<td>39330</td>
<td>623385</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1272</td>
<td>30300</td>
<td>696600</td>
<td>16418430</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2544</td>
<td>154500</td>
<td>6494400</td>
<td>241204950</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>309000</td>
<td>31534200</td>
<td>2068516800</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>63068400</td>
<td>9737380800</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>19474761600</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Computation becomes hard for large $n$ with the method of Cardona & Zhang which uses component graphs whose number increases rapidly!
$\text{TC}_{n,k}$ for small $n, k$ (ii)

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

\begin{align*}
\text{TC}_{n,k} &= n! 2^{n-1 - k} \omega_{n,k} \\
\omega_{n,k} &= \sum_{m \geq 1} b_{n,k,m} \\
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}.
\end{align*}
TC\(_{n,k}\) for small \(n, k\) (ii)

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

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

\[
TC_{n,k} = \frac{n!}{2^{n-1-k}}\omega_{n-1,k},
\]

where

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

with \(b_{n,k,m}\) given recursively by:

\[
b_{n,k,m} = \sum_{j=1}^{m} b_{n-1,k,j} + (n + m + k - 2) \sum_{j=1}^{m} b_{n-1,k-1,j}.
\]
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) \( \omega_i \) has either not occurred or it has occurred at least as often as \( \omega_j \) 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) \( \omega_i \) has either not occurred or it has occurred at least as often as \( \omega_j \) with \( j > i \).

For example, \( a_2 = 7 \) because

\[
\text{aaabbb, aababb, aabbab, ababab, ababab, baaabb, baabab.}
\]
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) \( \omega_i \) has either not occurred or it has occurred at least as often as \( \omega_j \) with \( j > i \).

For example, \( a_2 = 7 \) because

\[
\text{aaabbb, aababb, aabbab, ababab, ababab, baaabb, baabab.}
\]

Proposition (F., Yu, Zhang; 2021)

We have,

\[
\frac{TC_{n,n-1}}{n!} = a_{n-1}.
\]
Bijective Proof
Bijective Proof

Tree-Child Networks
Bijective Proof
Bijective Proof

Michael Fuchs (NCCU)

Tree-Child Networks

February 8th, 2023 9 / 15
Bijective Proof
Bijective Proof

\[ \rho \]

Michael Fuchs (NCCU)

Tree-Child Networks

February 8th, 2023
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):

\[ TC_n = \Theta(TC_{n+1} - \frac{2}{3} n^2 e^{\alpha_1(3n^{1/3})}) \]

where \( \alpha_1 \) 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,

\[ TC_n = \Theta(TC_{n,n-1}) = \Theta \left( n^{-2/3} e^{a_1(3n)^{1/3}} \left( \frac{12}{e^2} \right)^n n^{2n} \right), \]

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,

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

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

So, the base of the exponential growth rate is \( 12/e^{2!} \).
Chang & F. & Liu & Wallner & Yu (2023+) proved that

\[ \text{TC}_{n,n-1-k} \approx \frac{1}{2^k k!} \text{TC}_{n,n-1} \]

for \( k \) close to \( n \).
Chang & F. & Liu & Wallner & Yu (2023+) proved that

\[ \text{TC}_{n,n-1-k} \approx \frac{1}{2^k k!} \text{TC}_{n,n-1} \]

for \( k \) close to \( n \).

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

(a) We have,

\[ n - 1 - \# \text{ of reticulation nodes} \xrightarrow{d} \text{Poisson}(1/2). \]
Stochastic Results for TC-Networks

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

\[ \text{TC}_{n,n-1-k} \approx \frac{1}{2^k k!} \text{TC}_{n,n-1} \]

for \( k \) close to \( n \).

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

(a) We have,

\[ n - 1 - \# \text{ of reticulation nodes} \xrightarrow{d} \text{Poisson}(1/2). \]

(b) We have,

\[ \mathbb{E}(\# \text{ of cherries}) = \mathcal{O}(1). \]
General networks

Tree-sibling networks

Reticulation-visible networks

Tree-child networks

Galled networks

Normal networks

Galled trees

Rankable tree-child networks

Level-$k$ networks

Phylogenetic trees

Tree-Child Networks

February 8th, 2023
Results for Galled Networks

\( \text{GN}_n \) . . . \# of galled networks with \( n \) leaves.

Theorem (F., Yu, Zhang; 2022)

We have,

\[
\text{GN}_n \sim p^2 e^{4 \epsilon^4 n - 1/8 e^{2n/n^2} n^2}.
\]

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

\[
E(X_n) = n - 3/8 + o(1)
\]

and

\[
\text{Var}(X_n) = 3/4 + o(1).
\]
Results for Galled Networks

\( \text{GN}_n \ldots \# \) of galled networks with \( n \) leaves.

**Theorem (F., Yu, Zhang; 2022)**

We have,

\[
\text{GN}_n \sim \frac{\sqrt{2e} \sqrt[4]{e}}{4} n^{-1} \left( \frac{8}{e^2} \right)^n n^{2n}.
\]
Results for Galled Networks

\( \text{GN}_n \ldots \# \) of galled networks with \( n \) leaves.

**Theorem (F., Yu, Zhang; 2022)**

We have,

\[
\text{GN}_n \sim \frac{\sqrt{2e^4e}}{4} n^{1 - \frac{3}{8}} + o(1).
\]

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

\[
\mathbb{E}(X_n) = n - \frac{3}{8} + o(1) \quad \text{and} \quad \text{Var}(X_n) = \frac{3}{4} + o(1).
\]
Open Questions

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

\[(n - k)TC_{n,k} = (n + 1 - k)(n - k)TC_{n,k} - 1 + n(2n + k - 3)TC_{n-1,k}\]

Does there exist a constant \(\gamma\) such that

\[TC_n \sim \gamma n^{2/3} e^{1/(3n^{1/3})^{1/3}} e^{-n^{2/3}}\]

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


Enumeration and stochastic results for reticulation-visible networks?

Michael Fuchs (NCCU)
Open Questions

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

\[(n - k)TC_{n,k} = (n + 1 - k)(n - k)TC_{n,k-1} + n(2n + k - 3)TC_{n-1,k}\]
Open Questions

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

\[(n - k)TC_{n,k} = (n + 1 - k)(n - k)TC_{n,k-1} + n(2n + k - 3)TC_{n-1,k}\]

- Does there exist a constant \(\gamma\) such that

\[TC_n \sim \gamma n^{-2/3} e^{a_1(3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\]
Open Questions

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

\[(n - k)TC_{n,k} = (n + 1 - k)(n - k)TC_{n,k-1} + n(2n + k - 3)TC_{n-1,k}\]

- Does there exist a constant \(\gamma\) such that

\[TC_n \sim \gamma n^{-2/3} e^{a_1(3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\]

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

Open Questions

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

\[(n - k)TC_{n,k} = (n + 1 - k)(n - k)TC_{n,k-1} + n(2n + k - 3)TC_{n-1,k}\]

- Does there exist a constant \(\gamma\) such that

\[TC_n \sim \gamma n^{-2/3} e^{a_1(3n)^{1/3}} \left(\frac{12}{e^2}\right)^n n^{2n}\]

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


- Enumeration and stochastic results for reticulation-visible networks?
References


