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?

\[ X \ldots \text{a finite set.} \]
What is a (Binary) Phylogenetic Network?

$X$ ... a finite set.

Definition

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

(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?

\[ X \ldots \text{a finite set.} \]

**Definition**

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

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

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;

(b) *leaves*: in-degree 1 and out-degree 0; bijectively labeled by $X$;
What is a (Binary) Phylogenetic Network?

\[ X \ldots \text{a finite set.} \]

**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?

\( X \ldots \) 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.
A phylogenetic network is called tree-child network if every non-leaf node has at least one child which is not a reticulation node.
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) ![Diagram of a network that is not a tree-child network](image)

(b) ![Diagram of a network that is a tree-child network](image)
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:

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 $0 < c_1 < c_2$,*

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

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

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

For constants $0 < c_1 < c_2$,

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

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

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

Theorem (McDiarmid, Semple, Welsh; 2015)

For constants $0 < c_1 < c_2$,

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

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)

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

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

Cardona & Zhang (2020):

\[
\begin{array}{c|ccccccc}
  k \setminus n & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
  1 & 2 & 21 & 228 & 2805 & 39330 & 623385 \\
  2 & 42 & 1272 & 30300 & 696600 & 16418430 \\
  3 & 2544 & 154500 & 6494400 & 241204950 \\
  4 & 309000 & 31534200 & 2068516800 \\
  5 & 63068400 & 9737380800 \\
  6 & & & & & 19474761600 \\
\end{array}
\]

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 \( n, k \) (i)

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

Cardona & Zhang (2020):

\[
\begin{array}{c|ccccccc}
  k \backslash n & 2 & 3 & 4 & 5 & 6 & 7 \\
  \hline
  1 & 2 & 21 & 228 & 2805 & 39330 & 623385 \\
  2 & 42 & 1272 & 30300 & 696600 & 16418430 \\
  3 & 2544 & 154500 & 6494400 & 241204950 \\
  4 & 309000 & 31534200 & 2068516800 \\
  5 & 63068400 & 9737380800 \\
  6 & 19474761600 \\
\end{array}
\]

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

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

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

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

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

where

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

with $b_{n,k,m}$ given recursively by:

$$b_{n,k,m} = \sum_{j=1}^{m} b_{n-1,k,j} + (n + m + k - 2) \sum_{j=1}^{m} b_{n-1,k-1,j}.$$
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 $aaabbb, aababb, aabbab, abaabb, ababab, baabab$. 

**Proposition (F., Yu, Zhang; 2021)**

We have, $TC_n,n - 1 = n! = a_n - 1$. 

Michael Fuchs (NCCU) 
Tree-Child Networks 
February 8th, 2023 8 / 15
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

$$\text{aaabbb, aababb, aabbab, ababba, bababa, bbaabb, baabab}.$$
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

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

Proposition (F., Yu, Zhang; 2021)

We have,

$$\frac{\text{TC}_{n,n-1}}{n!} = a_{n-1}.$$
Bijective Proof
Bijective Proof
Bijective Proof
Bijective Proof
Bijective Proof

Tree-Child Networks

February 8th, 2023
Bijective Proof

Michael Fuchs (NCCU)
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):

Theorem (F., Yu, Zhang; 2021)

We have,

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

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,

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

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,

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

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

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 $n$. 

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

(a) We have, $n-1 - \# \text{ of reticulation nodes} \xrightarrow{d} \text{Poisson} \left(\frac{1}{2}\right)$.

(b) We have, $\mathbb{E}(\# \text{ of cherries}) = O(1)$. 

Michael Fuchs (NCCU)  Tree-Child Networks  February 8th, 2023  11 / 15
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,

\[ n - 1 - \# \text{ of reticulation nodes} \xrightarrow{d} \text{Poisson}(1/2). \]
Chang & F. & Liu & Wallner & Yu (2023+) proved that

$$TC_{n,n-1-k} \approx \frac{1}{2^k k!} TC_{n,n-1}$$

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,

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

Tree-sibling networks

Reticulation-visible networks

Tree-child networks

Galled networks

Normal networks

Galled trees

Rankable tree-child networks

Phylogenetic trees

Level-$k$ networks
Results for Galled Networks

$G\mathcal{N}_n \ldots \#$ of galled networks with $n$ leaves.
Results for Galled Networks

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

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

We have,

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

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

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

We have,

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

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

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

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

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

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

\[\text{TC}_n \sim \gamma n^{-2/3} e^{\alpha_1 (\frac{1}{2} n)}\]

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? E.g., for some results on the Sackin index see:

Open Questions

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

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


