Counting Phylogenetic Networks with the Component Graph Method

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

Michael Fuchs

Department of Mathematical Sciences
National Chengchi University

NATIONAL CHENGCHI UNIVERSITY

August 22nd, 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.

What is a (Binary) Phylogenetic Network?

X ... a finite set.

Definition

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

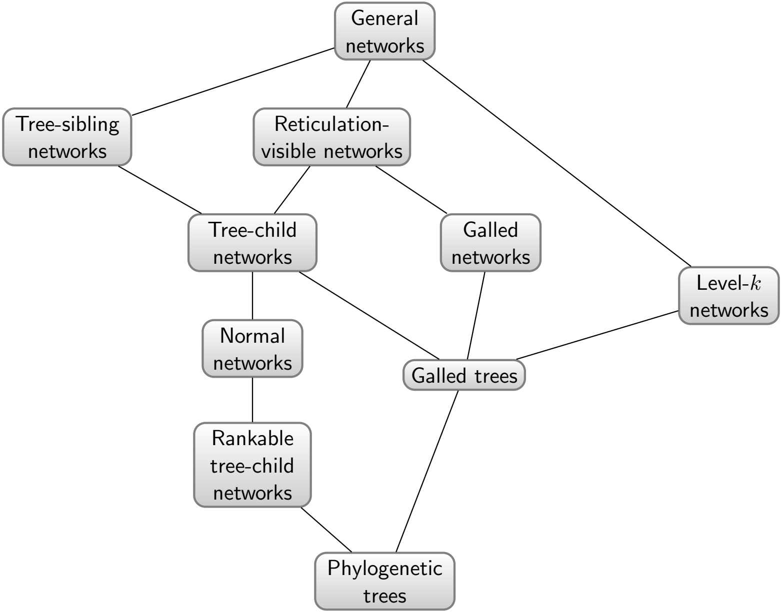
(a) *root: in-degree 0 and out-degree 1;*

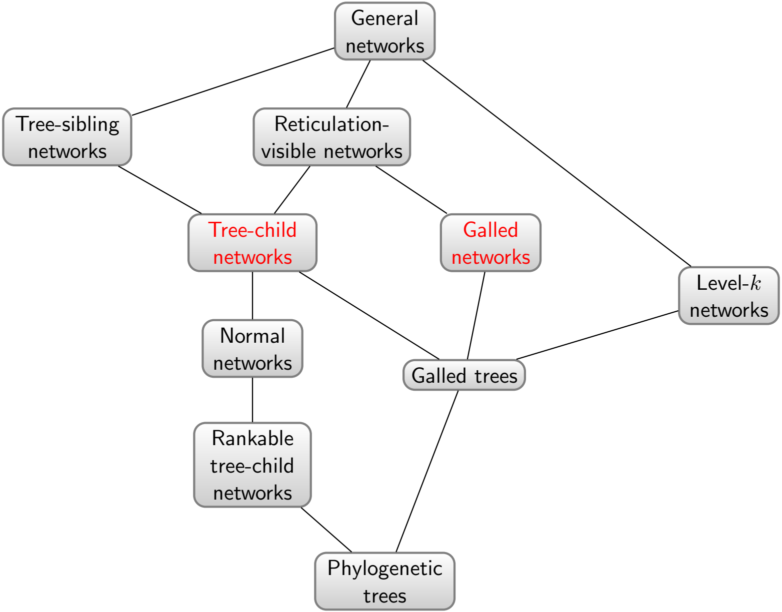
(b) *leaves: in-degree 1 and out-degree 0; bijectively labeled by X;*

(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.

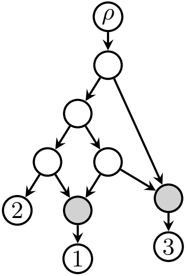
TC-Networks

Definition

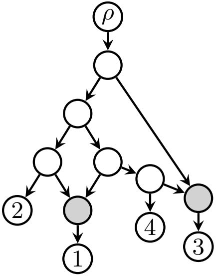
A phylogenetic network is called tree-child network if every non-leaf node has at least one child which is not a reticulation node.

Examples:

(a)



(b)



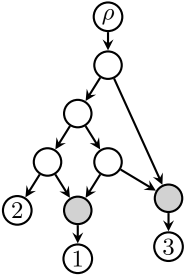
TC-Networks

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)

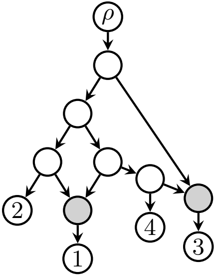


Figure: (a) is not a tc-network whereas (b) is a tc-network.

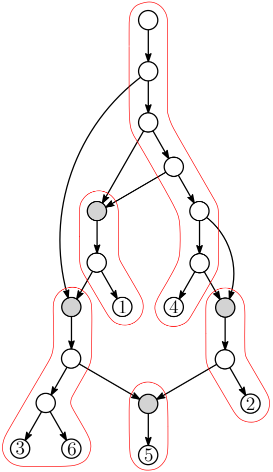
Method of Component Graphs

Cardona & Zhang (2020) used component graphs:

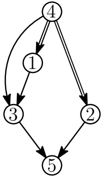
Method of Component Graphs

Cardona & Zhang (2020) used component graphs:

(a)



(b)



Counting TC-Networks

k_m ... # of component graphs with m nodes.

Proposition

k_m satisfies $k_m = \sum_{s=1}^{m-1} k_{m,s}$ where $k_{1,1} = 1$ and

$$k_{m,s} = \sum_{1\le t\le m-1-s} \binom{m}{s} \sum_{0\le \ell\le t} (-1)^{\ell} \binom{t}{\ell} \binom{m-s-\ell+1}{2}^{s} k_{m-s,t}.$$

Counting TC-Networks

k_m ... # of component graphs with m nodes.

Proposition

k_m satisfies $k_m = \sum_{s=1}^{m-1} k_{m,s}$ where $k_{1,1} = 1$ and

$$k_{m,s} = \sum_{1 \le t \le m-1-s} \binom{m}{s} \sum_{0 \le \ell \le t} (-1)^\ell \binom{t}{\ell} \binom{m-s-\ell+1}{2}^s k_{m-s,t}.$$

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

Theorem (Cardona & Zhang; 2020)

$$\mathrm{TC}_{n,k} = \frac{1}{2^{n-1-k}} \sum_{\{B_j\}_{j=1}^{k+1}} \sum_{G \in \mathcal{K}_{k+1}} \prod_{j=1}^{k+1} \frac{(2b_j + g_j - 2)!}{(b_j - 1)! \prod_{\ell=1}^{k+1} (g_{j,\ell})!}.$$

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

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

Lemma

In any tc-network: $k \leq n - 1$.

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

Lemma

In any tc-network: $k \le n-1$.

Cardona & Zhang:

$k \setminus n$	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

$TC_{n,k}$ for small n, k (i)

Lemma

In any tc-network: $k \leq n-1$.

Cardona & Zhang:

$k \setminus n$	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 more and more cumbersome because the number of component graphs 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:

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

Formulas for small k

Theorem (Cardona & Zhang; 2020)

We have,

$$\mathrm{TC}_{n,1} = \frac{n!(2n)!}{2^n n!} - 2^{n-1} n!.$$

and

$$\begin{aligned}\mathrm{TC}_{n,2} &= \frac{n!}{2^n} \sum_{j=1}^{n-2} \binom{2j}{j} \binom{2n-2j}{n-j} \frac{j(2j+1)(2n-j-1)}{2n-2j-1} \\ &\quad + n(n-1)n!2^{n-3} - \frac{(2n-1)!n}{3 \cdot 2^{n-1}(n-2)!} \\ &= n! \left(\frac{n(n+1)(n-1)(3n+2)}{6(2n+1)2^n} \binom{2n+2}{n+1} - n(n-1)2^n \right).\end{aligned}$$

Formulas for small k

Theorem (Cardona & Zhang; 2020)

We have,

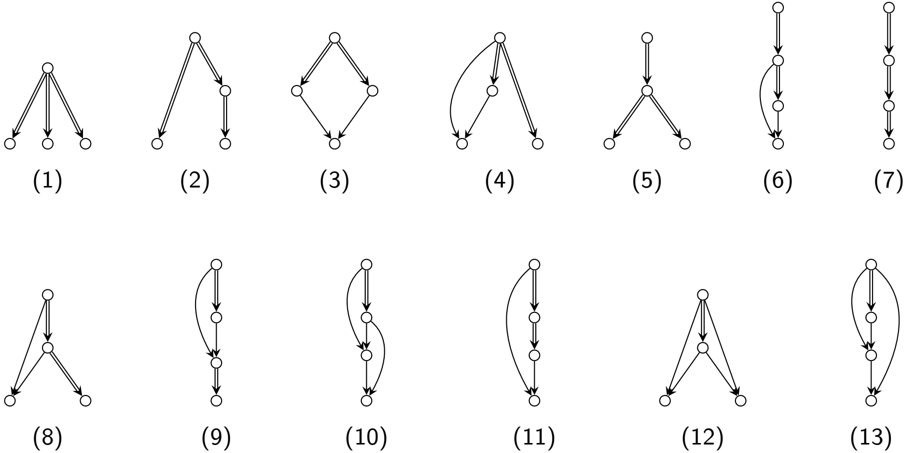
$$\mathrm{TC}_{n,1} = \frac{n!(2n)!}{2^n n!} - 2^{n-1}n!.$$

and

$$\mathrm{TC}_{n,2} = \frac{n!}{2^n}\sum_{j=1}^{n-2}\binom{2j}{j}\binom{2n-2j}{n-j}\frac{j(2j+1)(2n-j-1)}{2n-2j-1} + n(n-1)n!2^{n-3} - \frac{(2n-1)!n}{3\cdot 2^{n-1}(n-2)!}$$
$$= n!\left(\frac{n(n+1)(n-1)(3n+2)}{6(2n+1)2^n}\binom{2n+2}{n+1} - n(n-1)2^n\right).$$

En-Yu Huang (master student; 2022) derived a formula for $k = 3$.

Component Graphs for $k = 3$



Asymptotics of TC-Networks with fixed k

Proposition

Let $S_{n,k}$ be the number of tc-networks arising from the star-component graph. Then,

$$S_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^{n} n^{n+2k-1}.$$

Asymptotics of TC-Networks with fixed k

Proposition

Let $S_{n,k}$ be the number of tc-networks arising from the star-component graph. Then,

$$S_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^{n} n^{n+2k-1}.$$

In fact, as $n \to \infty$, $S_{n,k} \sim \mathrm{TC}_{n,k}$.

Asymptotics of TC-Networks with fixed k

Proposition

Let $S_{n,k}$ be the number of tc-networks arising from the star-component graph. Then,

$$S_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

In fact, as $n \to \infty$, $S_{n,k} \sim \mathrm{TC}_{n,k}$.

Theorem (F. & Huang & Yu; 2022)

As $n \to \infty$,

$$\mathrm{TC}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

Asymptotics of TC-Networks with fixed k

Proposition

Let $S_{n,k}$ be the number of tc-networks arising from the star-component graph. Then,

$$S_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

In fact, as $n \to \infty$, $S_{n,k} \sim \mathrm{TC}_{n,k}$.

Theorem (F. & Huang & Yu; 2022)

As $n \to \infty$,

$$\mathrm{TC}_{n,k} \sim \frac{2^{k-1}\sqrt{2}}{k!}\left(\frac{2}{e}\right)^n n^{n+2k-1}.$$

Galled Networks

Definition

A phylogenetic network is called a galled network if all its reticulation nodes are in a tree cycle.

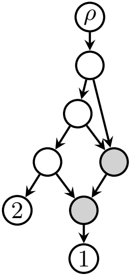
Galled Networks

Definition

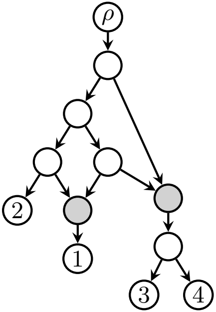
A phylogenetic network is called a galled network if all its reticulation nodes are in a tree cycle.

Examples:

(a)



(b)



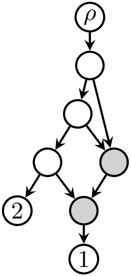
Galled Networks

Definition

A phylogenetic network is called a galled network if all its reticulation nodes are in a tree cycle.

Examples:

(a)



(b)

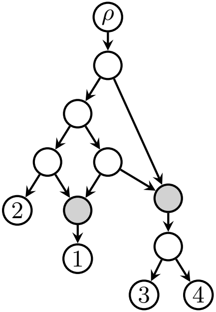
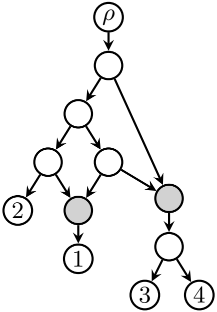


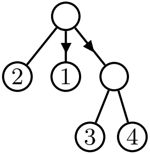
Figure: (a) is not a galled network whereas (b) is a galled network.

Component Graphs for Galled Networks

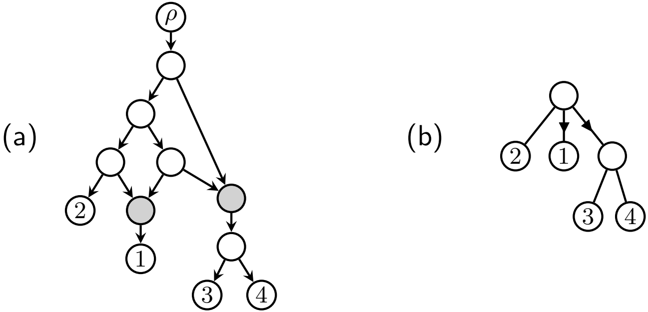
(a)



(b)



Component Graphs for Galled Networks



Theorem (Gunawan & Rathin & Zhang; 2022)

$$\mathrm{GN}_n = \sum_{\mathcal{T}} \prod_{v \in \mathcal{I}(\mathcal{T})} \sum_{j=c_{\mathrm{nlf}}(v)}^{c(v)} \binom{c_{\mathrm{lf}}(v)}{j - c_{\mathrm{nlf}}(v)} N_{c(v)+1}^{(j)}.$$

Asymptotics of Galled Networks (i)

We have,

$$\mathrm{OGN}_{n,k} = \binom{n}{k} N_{n+1}^{(k)},$$

where

$$N_n^{(k)} = (n+k-3)N_n^{(k-1)} + (k-1)N_n^{(k-2)} + \frac{1}{2}\sum_{1\le d\le k-1}\binom{k-1}{d}(2d-1)!!\left(N_{n-d}^{(k-1-d)} - N_{n-d+1}^{(k-1-d)}\right).$$

Asymptotics of Galled Networks (i)

We have,

$$\mathrm{OGN}_{n,k} = \binom{n}{k} N_{n+1}^{(k)},$$

where

$$\begin{aligned} N_n^{(k)} =& (n+k-3)N_n^{(k-1)} + (k-1)N_n^{(k-2)} \\ &+ \frac{1}{2}\sum_{1\le d\le k-1}\binom{k-1}{d}(2d-1)!!\left(N_{n-d}^{(k-1-d)} - N_{n-d+1}^{(k-1-d)}\right). \end{aligned}$$

Theorem (F. & Yu & Zhang; 2022)

As $n \to \infty$,

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

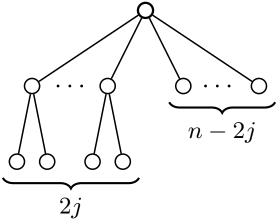
Asymptotics of Galled Networks (ii)

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

Asymptotics of Galled Networks (ii)

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

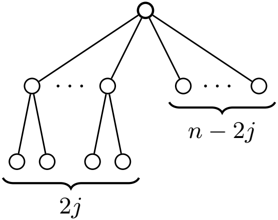
The following component graphs are dominating:



Asymptotics of Galled Networks (ii)

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

The following component graphs are dominating:



Theorem (F. & Yu & Zhang; 2022)

As $n \to \infty$,

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

Number of Reticulation Nodes

X_n ... number of reticulation nodes which are not followed by a leaf;

Y_n ... total number of reticulation nodes.

Number of Reticulation Nodes

X_n ... number of reticulation nodes which are not followed by a leaf;

Y_n ... total number of reticulation nodes.

Theorem (F. & Yu & Zhang; 2022)

We have,

$$(X_n, n - Y_n) \xrightarrow{d} (X, Y),$$

where for $j \geq 0$ *and* $k \geq -j$,

$$\mathbb{P}(X = j, Y = k) = \frac{e^{-7/8}}{16^j j!}[z^{j-k}]e^{1/(2z)}\left(1 + 2z + 3z^2\right)^j.$$

Number of Reticulation Nodes

X_n ... number of reticulation nodes which are not followed by a leaf;

Y_n ... total number of reticulation nodes.

Theorem (F. & Yu & Zhang; 2022)

We have,

$$(X_n, n - Y_n) \xrightarrow{d} (X, Y),$$

where for $j \geq 0$ *and* $k \geq -j$,

$$\mathbb{P}(X = j, Y = k) = \frac{e^{-7/8}}{16^j j!}[z^{j-k}]e^{1/(2z)}\left(1 + 2z + 3z^2\right)^j.$$

E.g., as a consequence,

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

Work in Progress and Open Problems

Work in Progress and Open Problems

- The component graph method can also be used to find formulas of the numbers of galled networks for small k and the first-order asymptotics of these numbers with fixed k.

Work in Progress and Open Problems

- The component graph method can also be used to find formulas of the numbers of galled networks for small k and the first-order asymptotics of these numbers with fixed k.

- The component graph of reticulation-visible networks is a tree-child network. This can be used to give a formula for the number of reticulation-visible networks with n leaves.

Work in Progress and Open Problems

- The component graph method can also be used to find formulas of the numbers of galled networks for small k and the first-order asymptotics of these numbers with fixed k.

- The component graph of reticulation-visible networks is a tree-child network. This can be used to give a formula for the number of reticulation-visible networks with n leaves.

- The formula for reticulation-visible networks can be used to give formulas for small k; it can also be used to obtain the first-order asymptotics for fixed k.

Work in Progress and Open Problems

- The component graph method can also be used to find formulas of the numbers of galled networks for small k and the first-order asymptotics of these numbers with fixed k.
- The component graph of reticulation-visible networks is a tree-child network. This can be used to give a formula for the number of reticulation-visible networks with n leaves.
- The formula for reticulation-visible networks can be used to give formulas for small k; it can also be used to obtain the first-order asymptotics for fixed k.
- What is the asymptotics of the number of reticulation-visible networks with n leaves? Does it contain a stretched example?

Some 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, 48 pages, submitted.

3. M. Fuchs, E.-Y. Huang, G.-R. Yu (2022). Counting phylogenetic networks with few reticulation vertices: a second approach, *Discrete Appl. Math.*, **320**, 140–149.

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. A. D. M. Gunawan, J. Rathin, L. Zhang (2020). Counting and enumerating galled networks, *Discrete Appl. Math.*, **283**, 644–654.