Counting Phylogenetic Networks

Charles Semple
School of Mathematics and Statistics
University of Canterbury, New Zealand

Joint work with Colin McDiarmid, Dominic Welsh

The Phylogenetic Network Workshop, Singapore, 2015

Questions

A phylogenetic network on X is a rooted acyclic directed graph with the following properties:

i. the root has out-degree two;
ii. vertices with out-degree zero have in-degree one (leaves), and the set of vertices with out-degree zero is X;
iii. all other vertices either have in-degree one and out-degree two (tree vertices), or in-degree two and out-degree one (reticulations).

- How many networks with leaf set X?
- Are there many more tree-child networks than normal networks?
- If one selects a network with leaf set X uniformly at random, what properties can one expect it to have when $|X|$ is sufficiently large?
  - Does it have a large number of reticulations?
  - What about the number of cherries?

Parameters of Networks

Theorem Let T be a binary phylogenetic tree on n vertices with m leaves. Then

$$n = 2m-1.$$

- The number of leaves bounds the total number of vertices.

Theorem Let N be a phylogenetic network on n vertices with m leaves, r reticulations, and t tree vertices. Then

$$m+r = t+2 = \frac{1}{2}(n+1).$$

- The total number of vertices is bounded either by the number of tree vertices or by the sum of the number of leaves and the number of reticulations.

Parameters of Tree-Child Networks

A network is tree-child if, for each non-leaf vertex, at least one of its children is a tree vertex or leaf.

Theorem Let N be a tree-child network with m leaves and r reticulations. Then

$$r \leq m-1.$$

Cardona, Rossello, Valiente (2009)

- For tree-child networks, the number of leaves bounds the total number of vertices.

Corollary Let N be a tree-child network on n vertices with m leaves and r reticulations. Then

$$r < \tfrac{1}{4}n < m.$$

McD, S, W (2015)

Counting Phylogenetic Trees

Theorem Let T_m denote the class of binary phylogenetic trees with leaf set [m]. Then

$$|T_m| = 1 \times 3 \times 5 \times \cdots \times (2m-3) = (2m-2)!/[(m-1)!\, 2^{m-1}]$$

Schröder (1870)

Corollary Let T_n denote the class of binary phylogenetic trees with vertex set [n]. Then

$$|T_n| = (n \text{ choose } m) \cdot (m-1)! \cdot |T_m| = (n \text{ choose } m)[(m-1)!/2^{m-1}]$$

Using Stirling's approximation,

$$|T_m| = 2^{m \log m + O(m)}$$

and

$$|T_n| = 2^{n \log n + O(n)}.$$

Counting Networks

Recall

$$|T_n| = 2^{n \log n + O(n)}.$$

Theorem Let GN_n denote the class of (general) networks with vertex set $[n]$. Then

$$|GN_n| = 2^{(3/2)n \log n + O(n)}.$$

Equivalently, there exists positive integers c_1 and c_2 such that

$$(c_1 n)^{(3/2)n} \leq |GN_n| \leq (c_2 n)^{(3/2)n}.$$

McD, S, W (2015)

Proof (Upper Bound)

- Find an upper bound for the number $f(n, m)$ of (simple, undirected) graphs on vertex set $[n]$ with m vertices of degree 1, one vertex of degree 2, and remaining vertices of degree 3.
- Use a configuration model with $3n - 2m - 1$ labelled points partitioned into $m + 1 + (n-m-1)$ parts.
- Number of perfect matchings is

$$(3n - 2m - 2)!! \leq (3n)^{(3/2)n-m}.$$

- Therefore

$$f(n, m) \leq n \cdot (n \text{ choose } m) \cdot (3n)^{(3/2)n-m}.$$

- Thus the number $g(n, m)$ of networks in GN_n with m leaves is

$$g(n, m) \leq 2^{3n} \cdot n \cdot 2^n \cdot (3n)^{(3/2)n-m}.$$

  So

$$g(n, m) \leq d^n n^{(3/2)n-m+1}$$

  for some constant d.
- Summing over $m \geq 1$, for some constant c,

$$|GN_n| \leq c^n n^{(3/2)n}.$$

Proof (Lower Bound)

- Let G be a cubic graph on $[n]$.
- Suppose G has a Hamiltonian cycle $C = v_1 v_2 \cdots v_n v_1$.
- Orient G by directing each edge $\{v_i, v_j\}$ from v_i to v_j if $i < j$.
- Construct a network by deleting (v_1, v_n), and adding new vertices p, m_1, m_2, and new edges (p, v_1), (p, m_1), and (v_n, m_2).
- Each cubic graph on $[n]$ with a Hamiltonian cycle yields a distinct network.
- For all sufficiently large n, the number of cubic graphs on $[n]$ is at least $d^n n^{(3/2)n}$ for some constant d.
- Almost all cubic graphs on $[n]$ are Hamiltonian (Robinson, Wormald 1992).
- Hence, for some constant c,

$$|GN_n| \geq c^n n^{(3/2)n}.$$

Counting Tree-Child and Normal Networks

Recall $|T_n| = 2^{n \log n + O(n)}$ and $|T_m| = 2^{m \log m + O(m)}$.

A tree-child network is normal if it has no short cuts.

Theorem Let NL_n and TC_n denote the classes of normal and tree-child networks with vertex set $[n]$. Then

$$|NL_n| = 2^{(5/4)n \log n + O(n)}$$

and

$$|TC_n| = 2^{(5/4)n \log n + O(n)}.$$

McD, S , W (2015)

Theorem Let NL_m and TC_m denote the classes of normal and tree-child networks with leaf set $[m]$. Then

$$|NL_m| = 2^{2m \log m + O(m)}$$

and

$$|TC_m| = 2^{2m \log m + O(m)}.$$

McD, S, W (2015)

Almost All Tree-Child Networks

Almost all networks in TC_n have some property if the proportion of networks in TC_n with the property tends to 1 as n tends to ∞.

Theorem

i. Almost all networks in TC_n are not normal.
ii. Almost all networks in TC_m are not normal.

McD, S, W (2015)

Almost All Networks

Theorem

i. Almost all networks in GN_n have $o(n)$ leaves and $(\frac{1}{2} + o(1))n$ reticulations.
ii. Almost all networks in TC_n have $(\frac{1}{4} + o(1))n$ leaves and $(\frac{1}{4} + o(1))n$ reticulations.
iii. Almost all networks in TC_m have $(1 + o(1))m$ reticulations and $(4 + o(1))m$ vertices in total.

McD, S, W (2015)

A twig is a non-leaf vertex in a pendant subtree.

Theorem

i. Almost all networks in TC_n have $o(n)$ twigs.
ii. Almost all networks in TC_m have $o(m)$ twigs.

McD, S, W (2015)

- Almost all n-vertex tree-child networks have $n/4$ leaves but only $o(n)$ twigs.

Final Remarks

- Almost all networks in GN_n have at most $O(n/\log n)$ leaves.
  - Is this the right order of magnitude or is there far fewer leaves?
- The depth of a network is the maximum length of a directed path from the root to a leaf.
  - The depth of an n-vertex network is at least $\log n - 1$.
  - Our constructions suggest that typical normal and tree-child networks have small depth, and typical general networks have much greater depth.
  - How large are these typical depths?