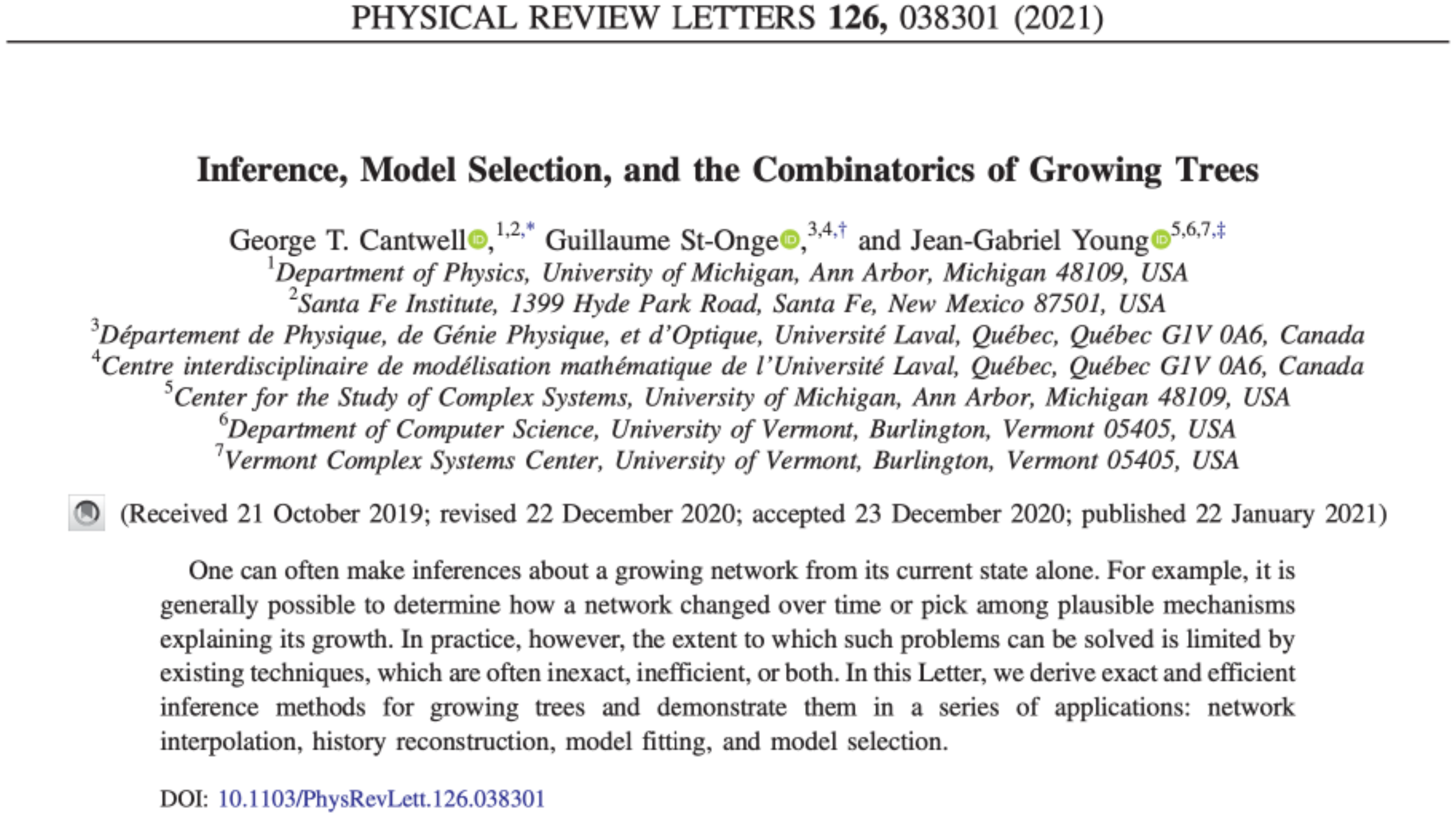
Inference on growing trees

Kevin Ly
AIG 5/10/21



PHYSICAL REVIEW LETTERS **126,** 038301 (2021)

Inference, Model Selection, and the Combinatorics of Growing Trees

George T. Cantwell,[1,2,*] Guillaume St-Onge,[3,4,†] and Jean-Gabriel Young[5,6,7,‡]
[1] *Department of Physics, University of Michigan, Ann Arbor, Michigan 48109, USA*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA*
[3] *Département de Physique, de Génie Physique, et d'Optique, Université Laval, Québec, Québec G1V 0A6, Canada*
[4] *Centre interdisciplinaire de modélisation mathématique de l'Université Laval, Québec, Québec G1V 0A6, Canada*
[5] *Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA*
[6] *Department of Computer Science, University of Vermont, Burlington, Vermont 05405, USA*
[7] *Vermont Complex Systems Center, University of Vermont, Burlington, Vermont 05405, USA*

(Received 21 October 2019; revised 22 December 2020; accepted 23 December 2020; published 22 January 2021)

One can often make inferences about a growing network from its current state alone. For example, it is generally possible to determine how a network changed over time or pick among plausible mechanisms explaining its growth. In practice, however, the extent to which such problems can be solved is limited by existing techniques, which are often inexact, inefficient, or both. In this Letter, we derive exact and efficient inference methods for growing trees and demonstrate them in a series of applications: network interpolation, history reconstruction, model fitting, and model selection.

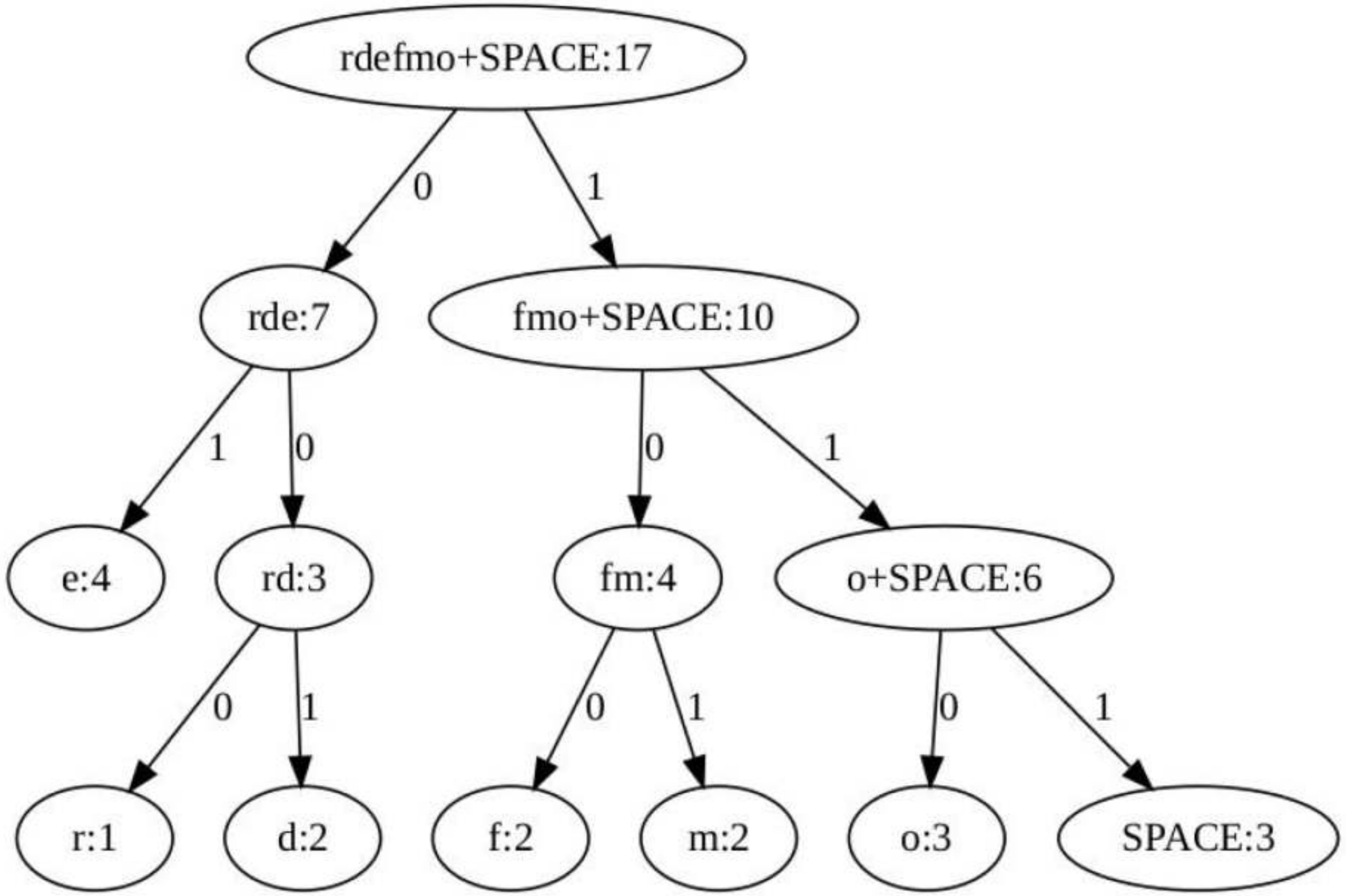
DOI: 10.1103/PhysRevLett.126.038301

Growing trees

- Trees: simple undirected, acyclic graph. If there are n nodes, then there are $n-1$ edges

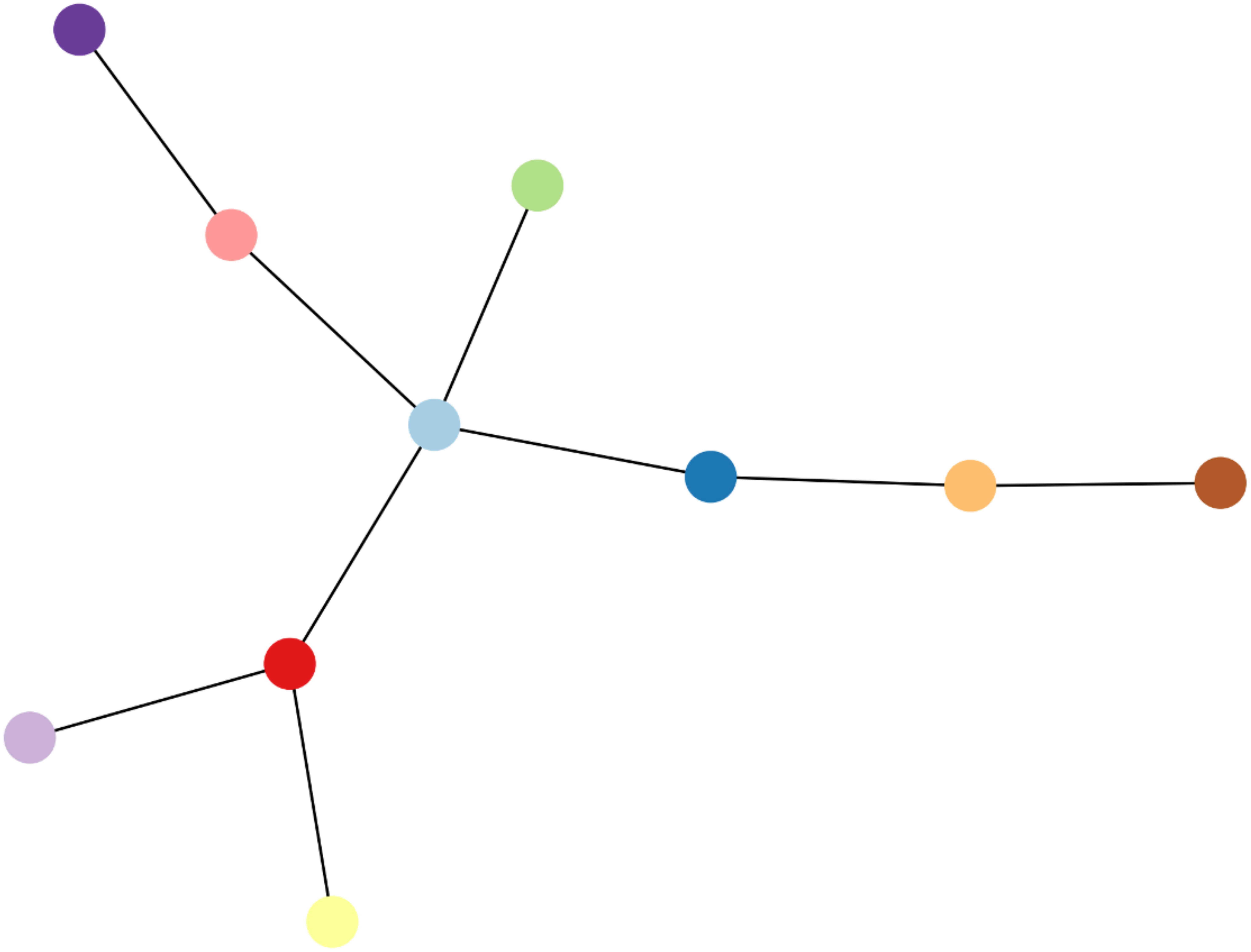
- Growing: each node is added one at a time, such that the "history" of a growing tree can be represented as a sequence of nodes

- Examples: Academic trees, social networks

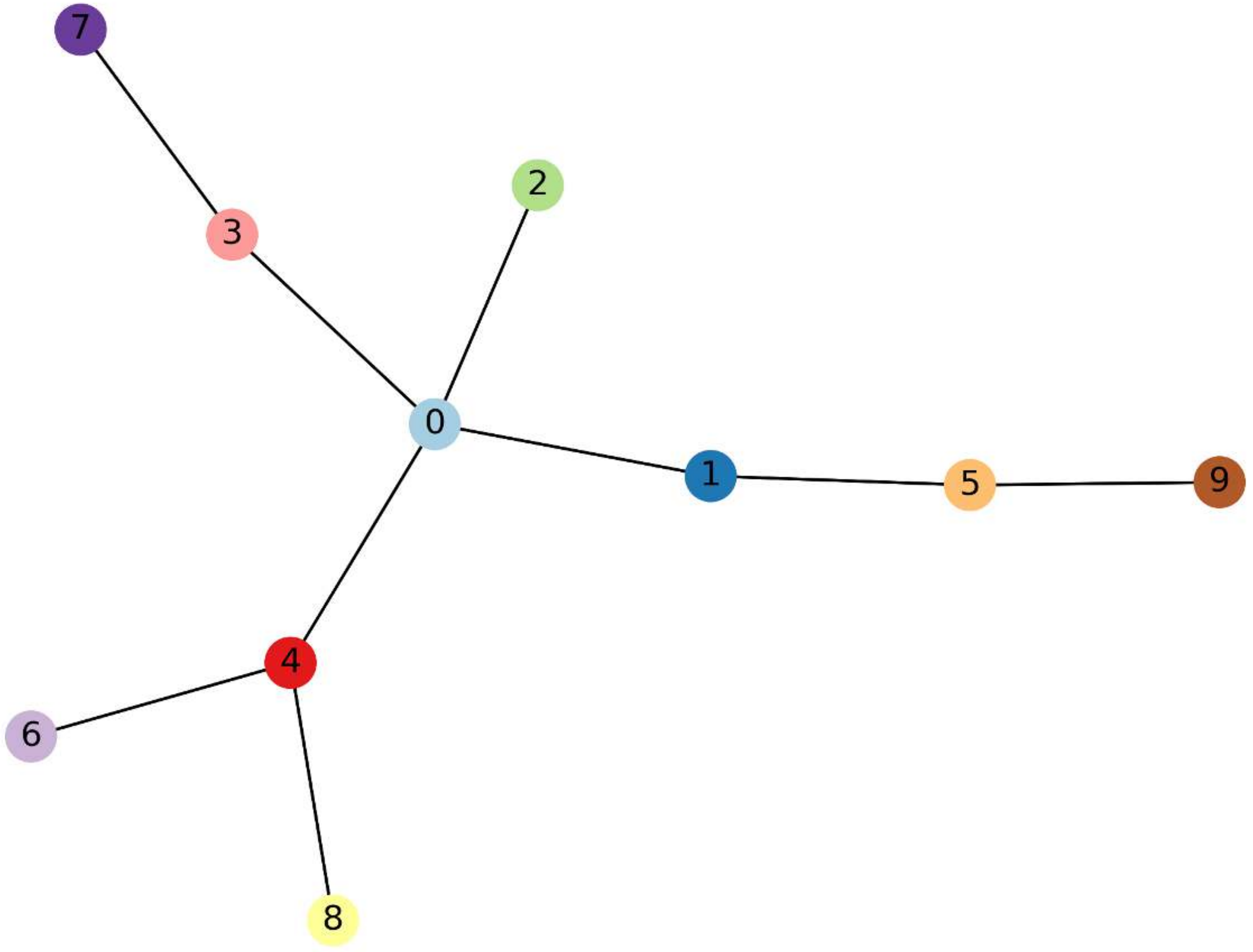


A tree (not a growing one), from Chad's talk on Huffman Encoding

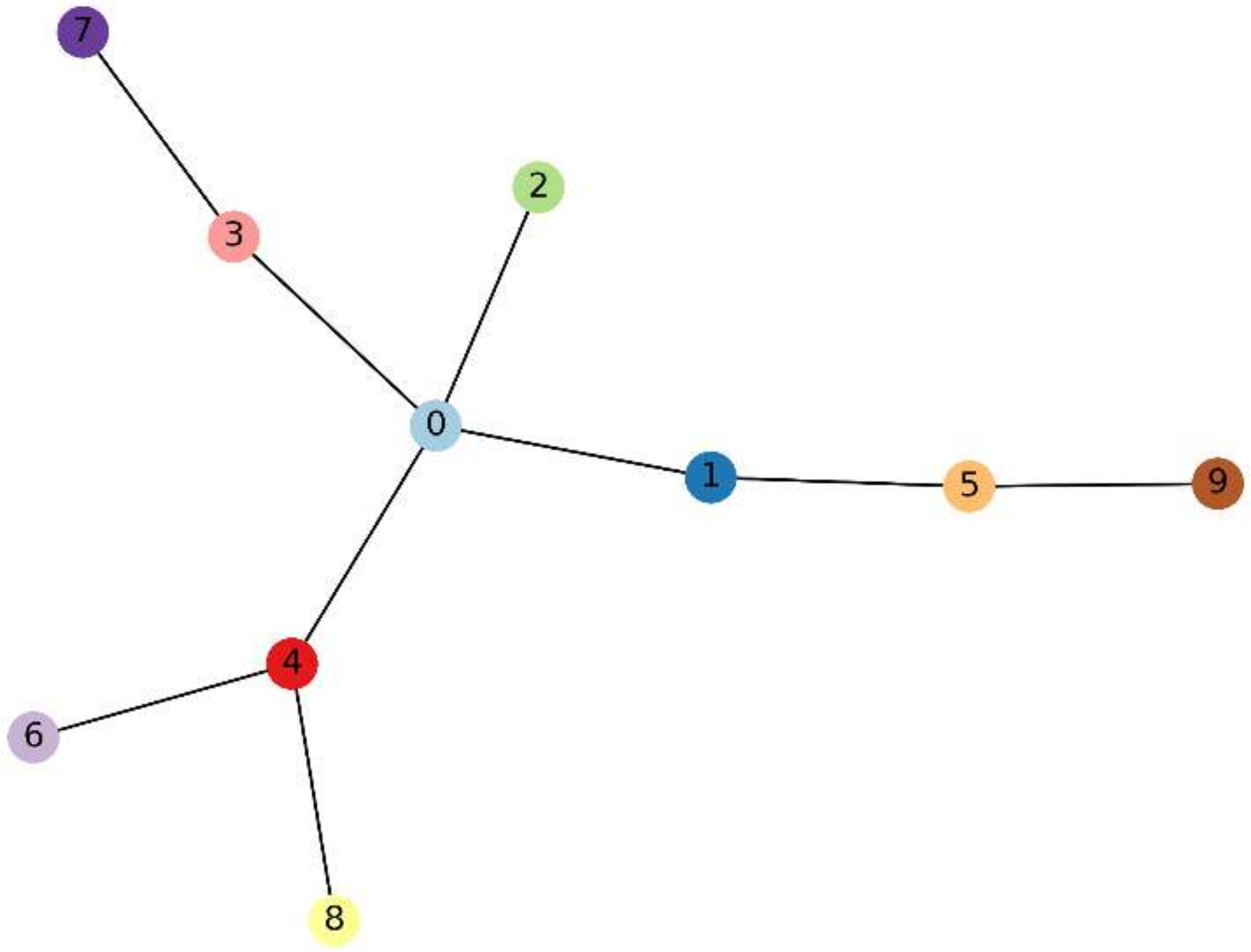
Which node came first?

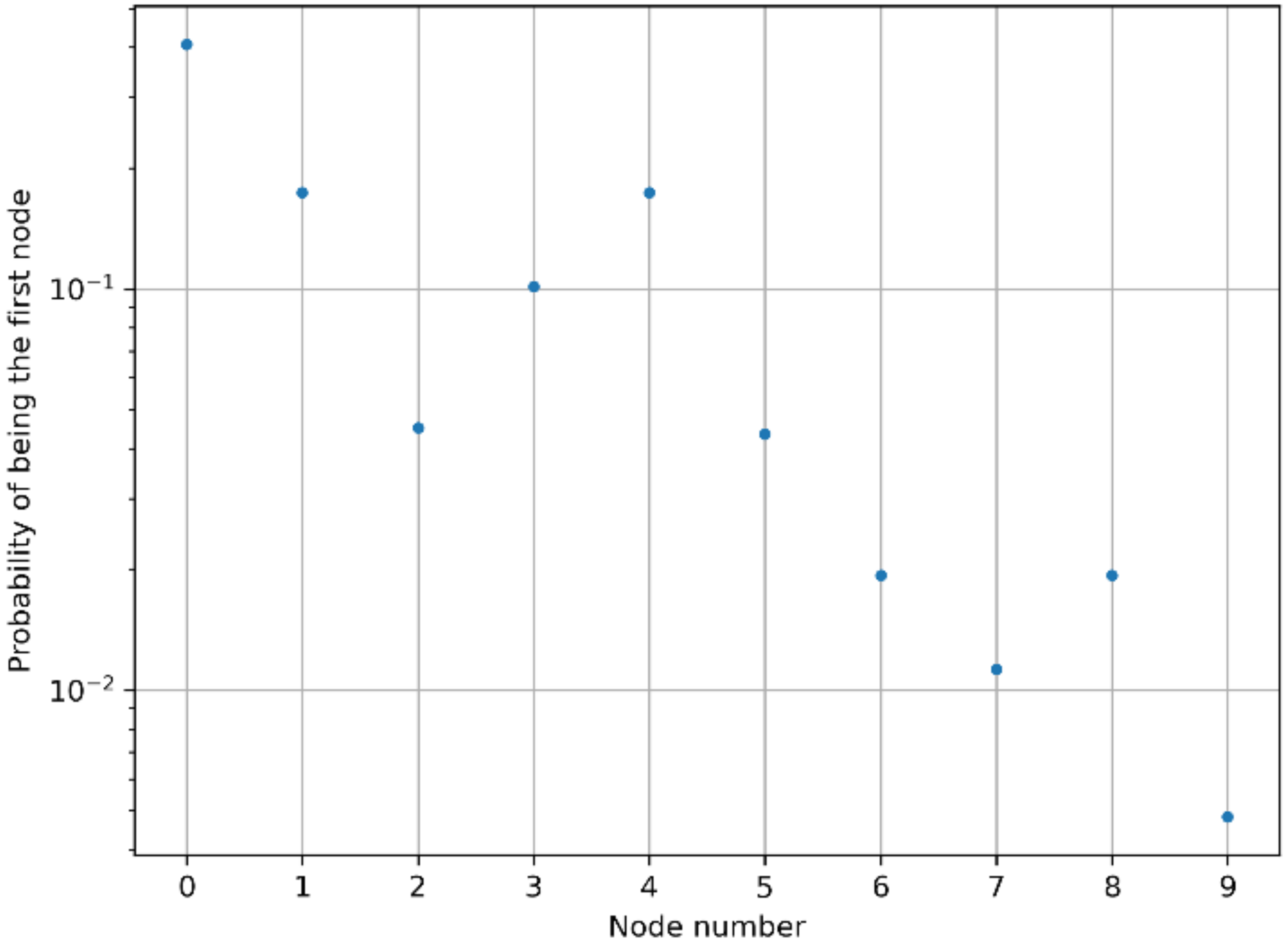


Which node came first?



Which node came first?





Histories

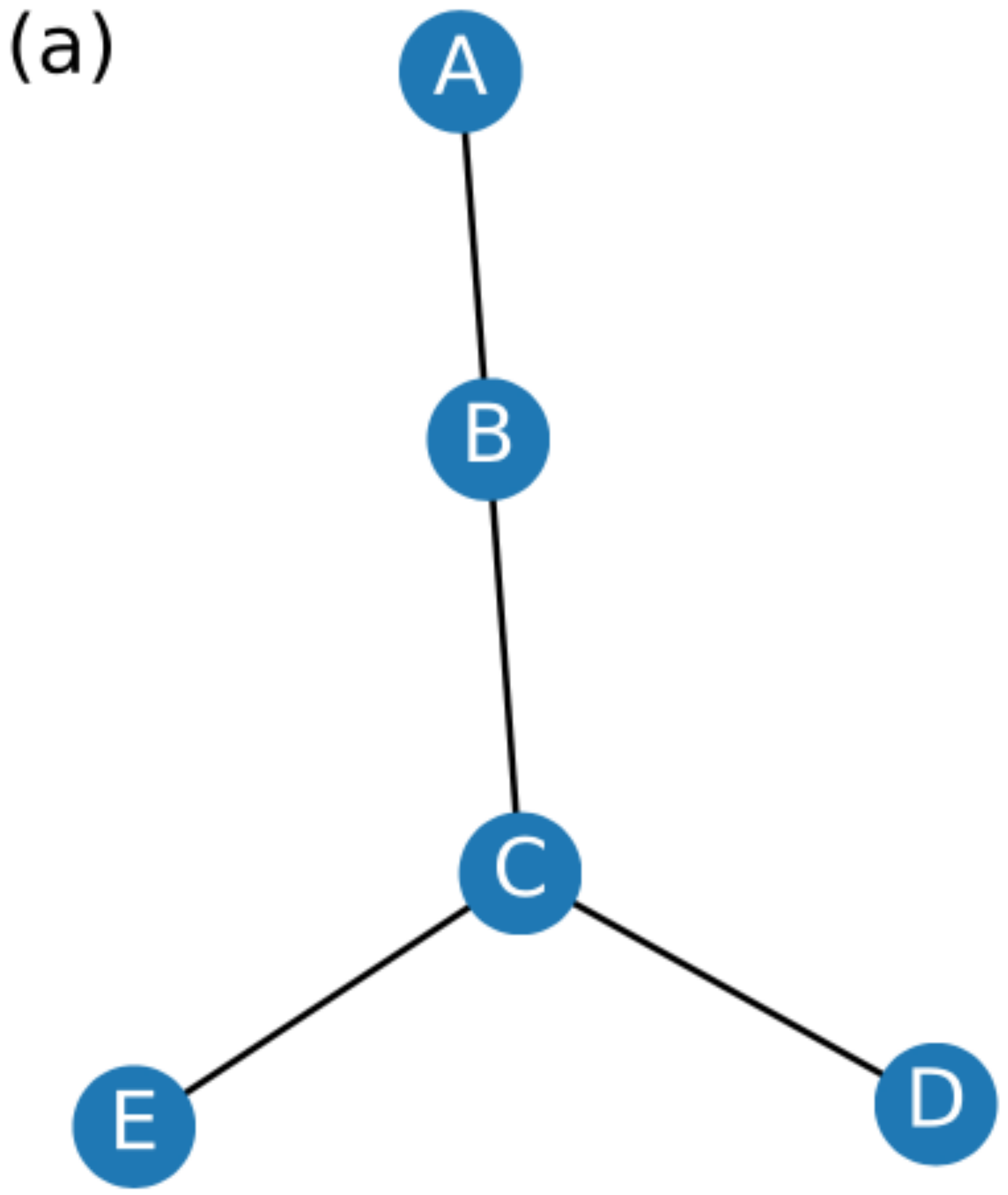


TABLE I. Enumeration of all consistent histories for the network of Fig. 1.

ABCDE	ABCED	BACDE	BACED
BCADE	BCAED	BCDAE	BCDEA
BCEAD	BCEDA	CBADE	CBAED
CBDAE	CBDEA	CBEAD	CBEDA
CDBAE	CDBEA	CDEBA	CEBAD
CEBDA	CEDBA	DCBAE	DCBEA
DCEBA	ECBAD	ECBDA	ECDBA

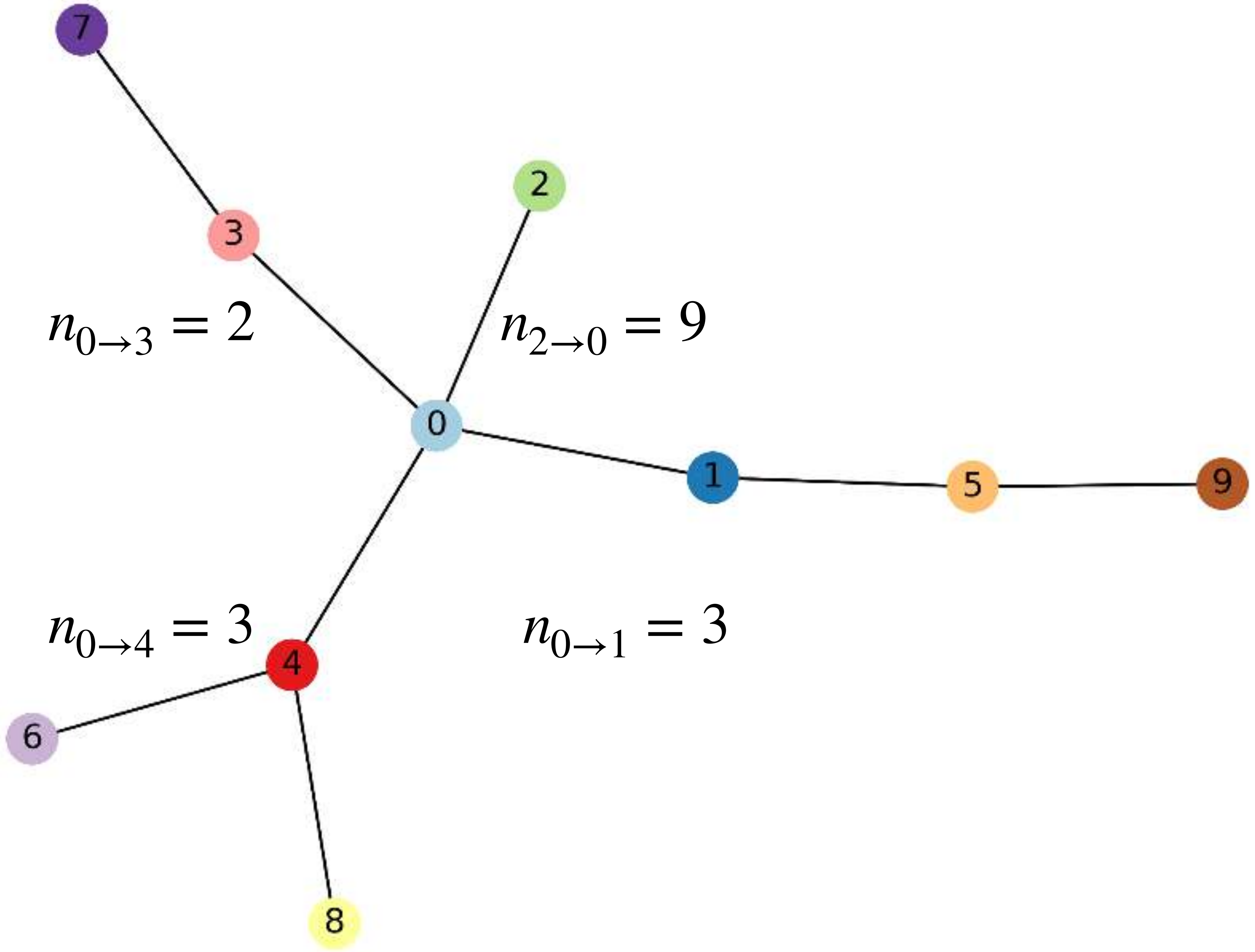
Of the 120 possible sequences of nodes, only 28 are possible histories.
For example, the sequence EABCD is not possible

Counting descendants

Let $n_{i\to j}$ be the number of nodes in the branch containing node j when edge (i, j) is removed. Then

$$n_{i\to j} = 1 + \sum_{k\in N_j|i} n_{j\to k}$$

where the sum is over all of j's neighbors except for i

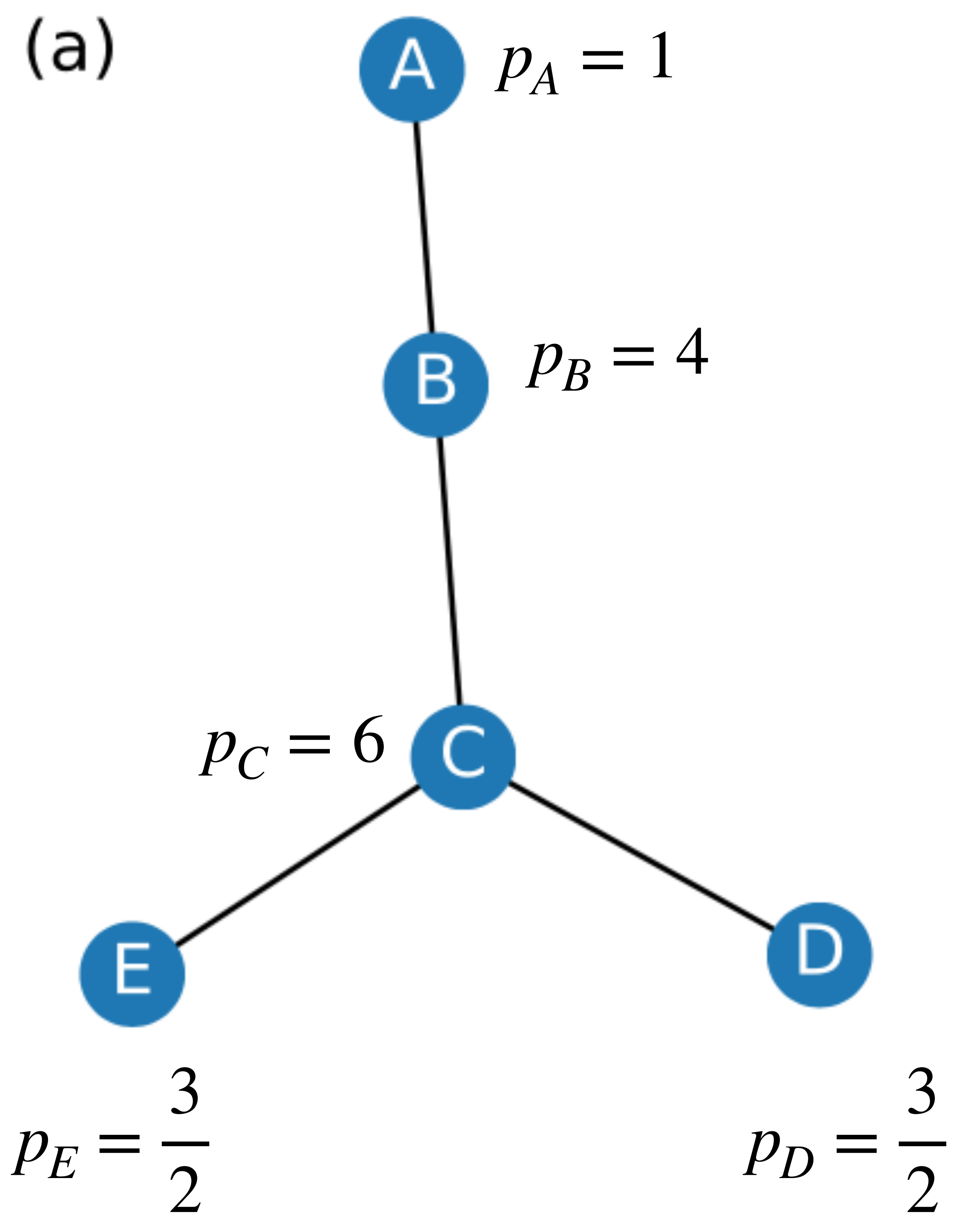


Probability of being the root

Let p_i be the probability that node i comes first. Set $p_0 = 1$ (arbitrary starting point), then

$$p_j = p_i \left(\frac{n_{i \to j}}{n - n_{i \to j}} \right)$$

will give the unnormalized probabilities (will be proven momentarily)



example starting at node A

Probability of being the root

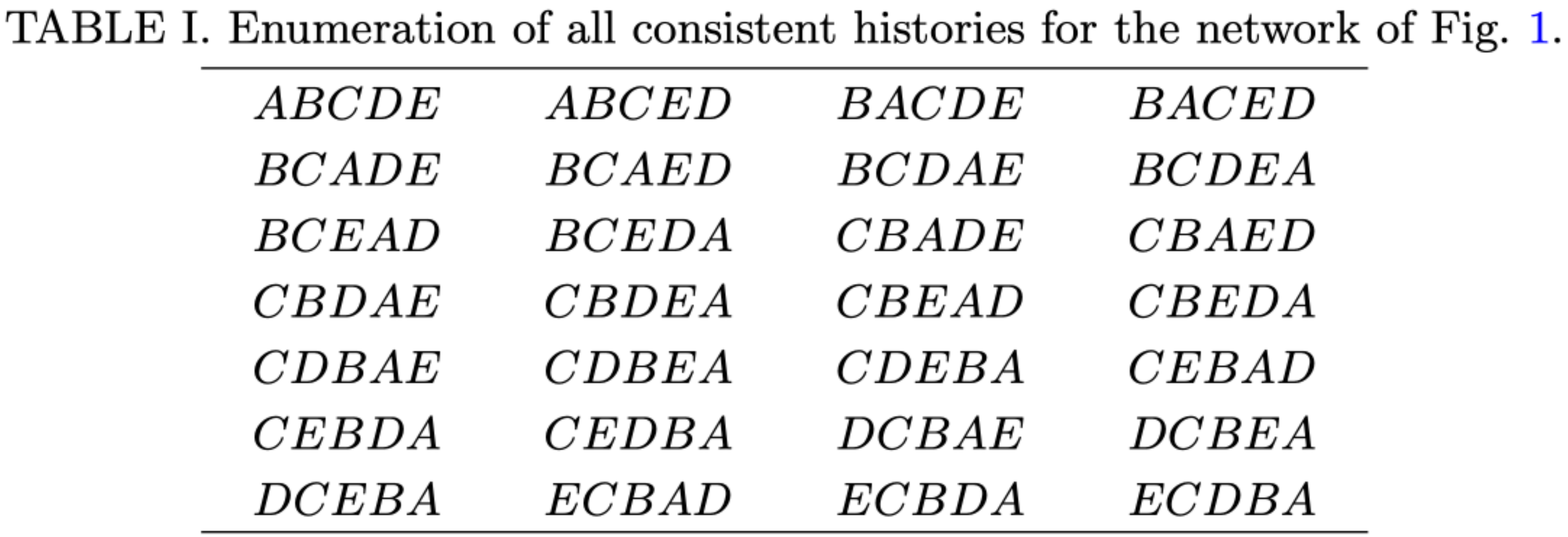
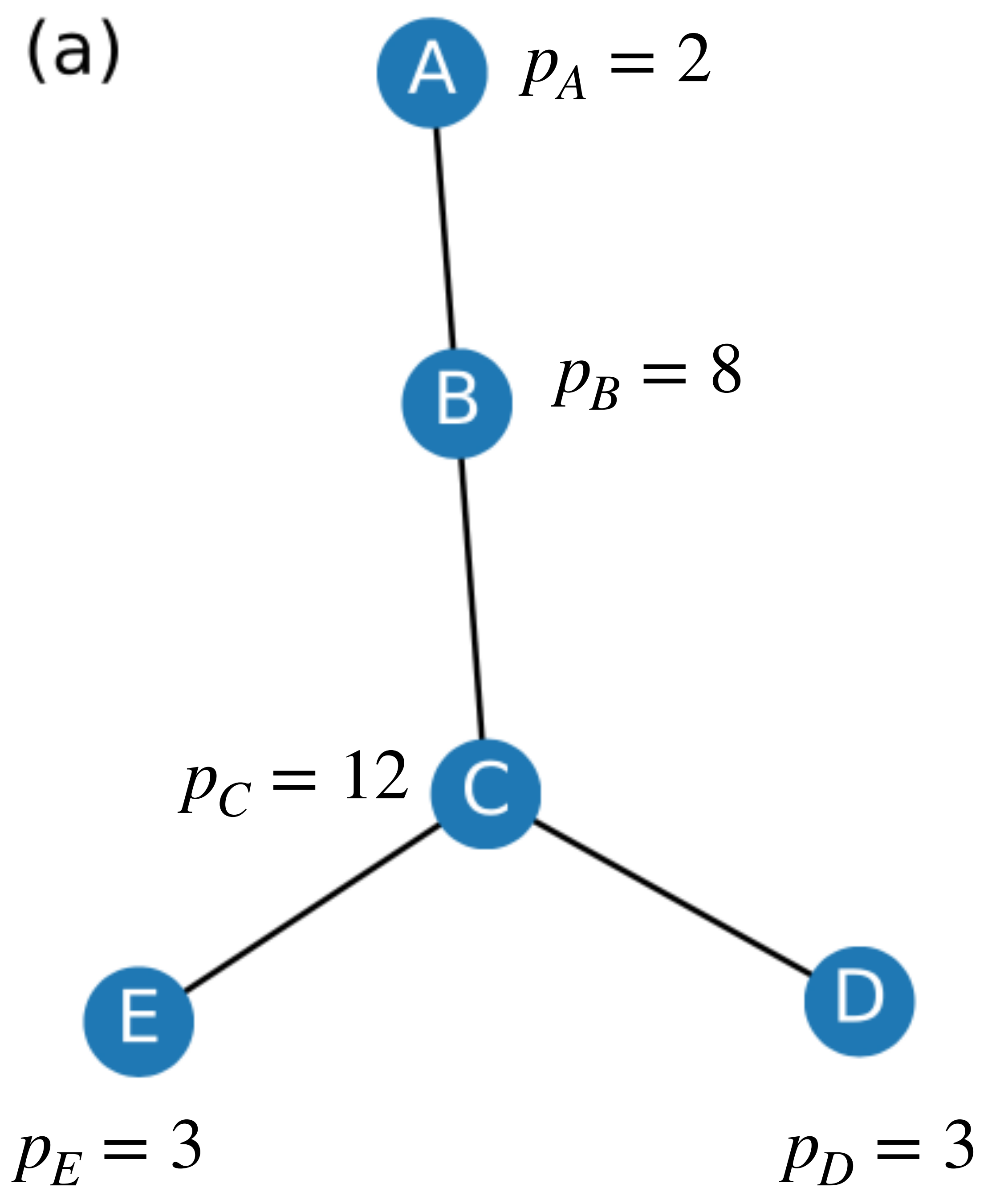


TABLE I. Enumeration of all consistent histories for the network of Fig. 1.

$ABCDE$	$ABCED$	$BACDE$	$BACED$
$BCADE$	$BCAED$	$BCDAE$	$BCDEA$
$BCEAD$	$BCEDA$	$CBADE$	$CBAED$
$CBDAE$	$CBDEA$	$CBEAD$	$CBEDA$
$CDBAE$	$CDBEA$	$CDEBA$	$CEBAD$
$CEBDA$	$CEDBA$	$DCBAE$	$DCBEA$
$DCEBA$	$ECBAD$	$ECBDA$	$ECDBA$



multiplying all the previous results by 2,
we see this is the right answer!

Counting histories

i.e. $p_i = \frac{h_i}{\sum_j h_j}$

Let h_i be the number of histories in which i comes first, and $h_{i\to j}$ be the number of histories of the subtree rooted at j when edge (i,j) is removed. Then

$$h_i = (n-1)!\prod_{j\in N_i}\frac{h_{i\to j}}{n_{i\to j}!}$$

NOT $\prod_{j\in N_i} h_{i\to j}$!!!!!!!

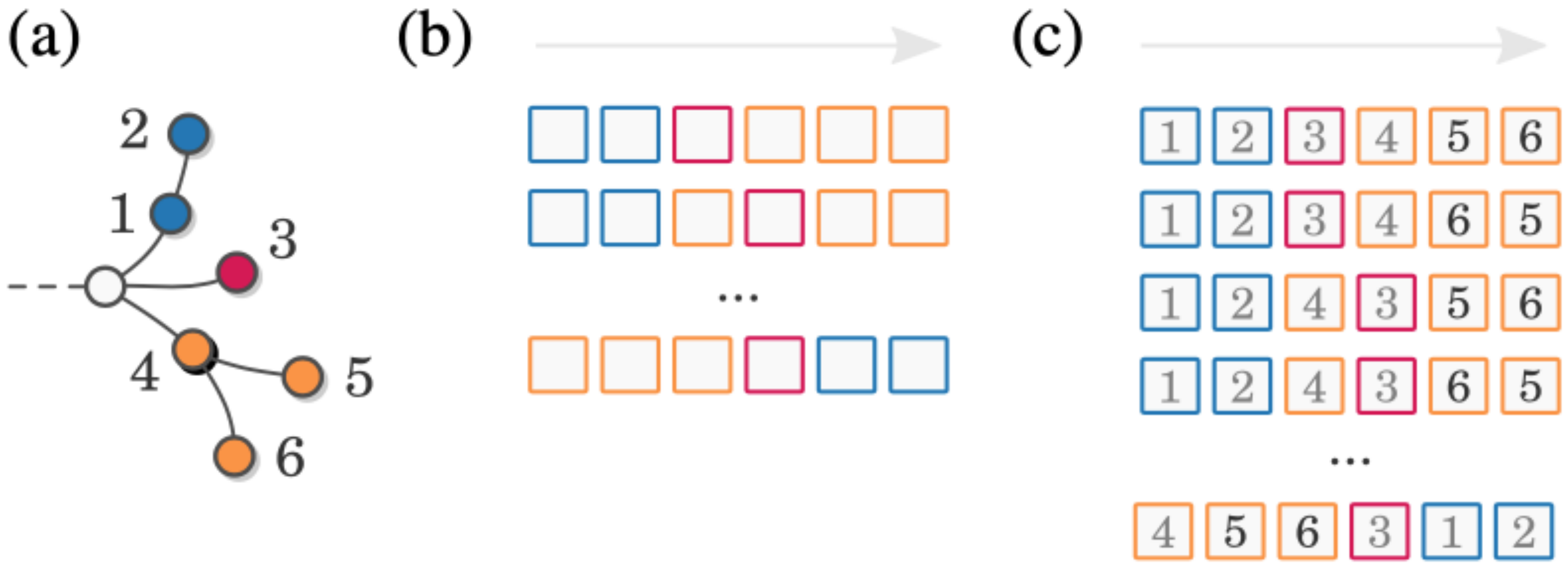


FIG. 1. Interlacing of histories. (a) We count the number of histories for the tree rooted on the focal node, shown in white. (b) There are $(6!/1!2!3!) = 60$ ways to interlace the nodes of the three subtrees, and (c) $1\times 1\times 2 = 2$ distinct orderings of these nodes for each of the interlacings. The total number of possible histories for the subtree rooted on the white node is therefore equal to $2\times 60 = 120$.

Proof of $p_j = p_i(\ldots)$

Note that the same game played in previous slide works for subtrees:

$$h_i = (n-1)!\prod_{j\in N_i}\frac{h_{i\to j}}{n_{i\to j}!}$$

$$\implies h_{i\to j} = (n_{i\to j}-1)!\prod_{k\in N_j|i}\frac{h_{j\to k}}{n_{j\to k}!}$$

$$\implies h_{i\to j}\frac{h_{j\to i}}{n_{j\to i}!}\frac{(n-1)!}{(n_{i\to j}-1)!} = h_j$$

Proof of $p_j = p_i(\ldots)$

$$h_j = h_{i\to j}\frac{h_{j\to i}}{n_{j\to i}!}\frac{(n-1)!}{(n_{i\to j}-1)!}$$

$$\implies \frac{p_j}{p_i} = \frac{h_j}{h_i} = \frac{n_{i\to j}!(n_{j\to i}-1)!}{n_{j\to i}!(n_{i\to j}-1)!} = \frac{n_{i\to j}}{n_{j\to i}} = \frac{n_{i\to j}}{n-n_{i\to j}}$$

noting that $n_{i\to j} + n_{j\to i} = n$. This is the relation we used to calculate all of the root probabilities

Probability of arriving at t

Let $p_i(t)$ be the probability that node i arrived at time t, and $g_{i\to j}(t)$ be the number of histories in which node j arrived *before* time t in the subtree containing j when edge (i, j) is removed. Then

$$p_i(t) = \frac{1}{\underline{Z}} \sum_{j \in N_i} \underline{g_{i\to j}(t)}\, \underline{h_{j\to i}} \underline{\binom{n-t-1}{n_{i\to j}-t}}$$

Normalization, number of histories in which j comes before t excluding any histories containing i and its descendants, number of histories of the subtree rooted at i, number of subsequences of both subtrees consistent with previous two occurrences

Probability of arriving at t

Let $h_{i,k \to j}$ be the number of histories rooted at j when the edges (i, j) and (k, j) are removed:

$$h_{i,k \to j} = (n_{k \to j} - 1 - n_{j \to 1})! \prod_{l \in N_j | i,k} \frac{h_{j \to l}}{n_{j \to l}!}$$

similar to previous expression for counting histories but minus another branch

$$\implies g_{i \to j}(t+1) - g_{i \to j}(t) = \sum_{j \in N_j | i} g_{j \to k}(t) h_{i,k \to j} \binom{n - t - 1}{n_{j \to k} - t}$$

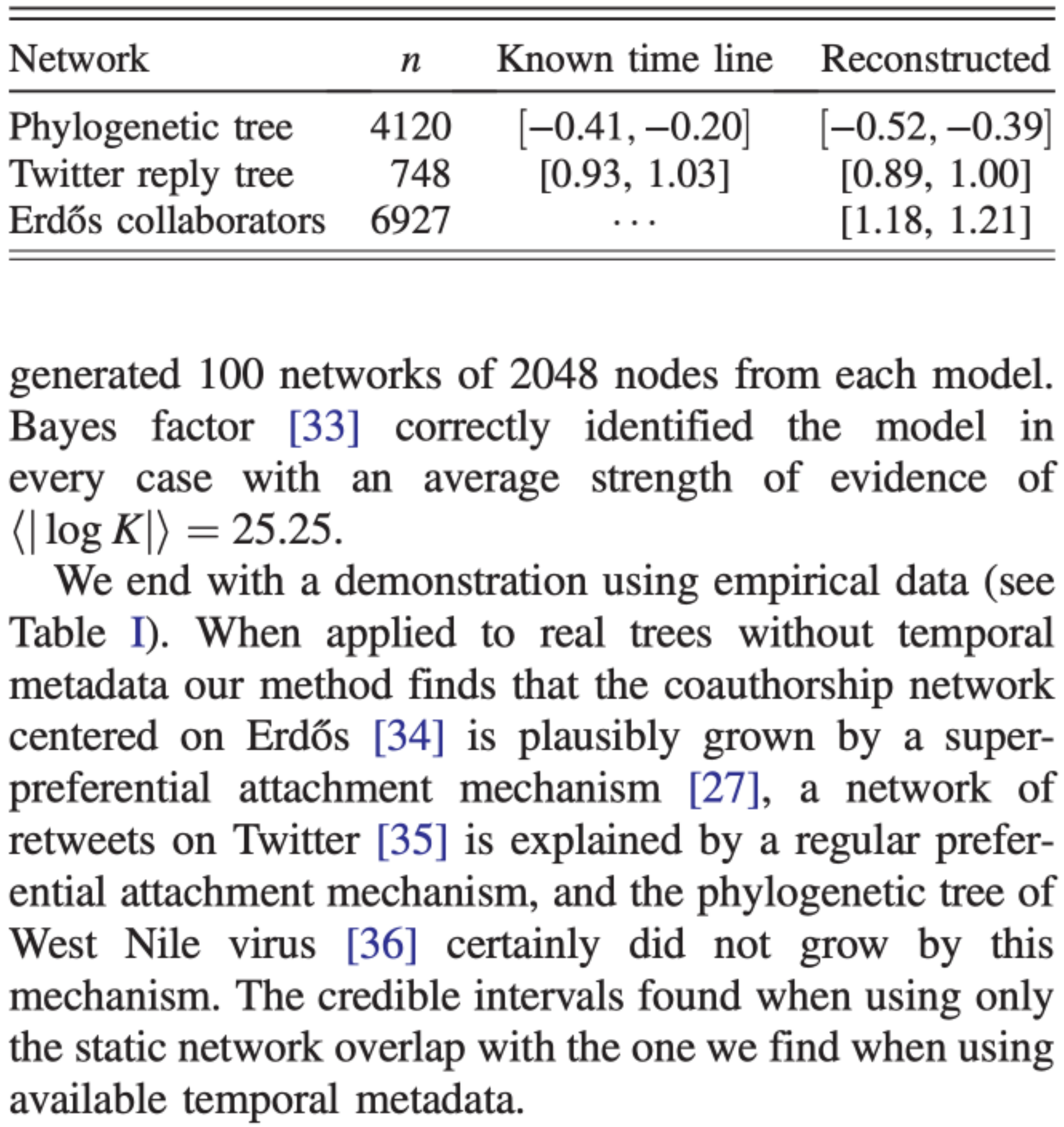
Demo

Summary

- Presented a means of counting histories and calculating probabilities for growing trees

- $p_i(t)$ calculation is $\mathcal{O}(n^2)$, whereas the number of possible (but not necessarily consistent) sequences of nodes is $n!$

- Probabilities can be used to do model selection on growing trees



Network	n	Known time line	Reconstructed
Phylogenetic tree	4120	$[-0.41, -0.20]$	$[-0.52, -0.39]$
Twitter reply tree	748	$[0.93, 1.03]$	$[0.89, 1.00]$
Erdős collaborators	6927	$\cdots$	$[1.18, 1.21]$

generated 100 networks of 2048 nodes from each model. Bayes factor [33] correctly identified the model in every case with an average strength of evidence of $\langle |\log K| \rangle = 25.25$.

We end with a demonstration using empirical data (see Table I). When applied to real trees without temporal metadata our method finds that the coauthorship network centered on Erdős [34] is plausibly grown by a superpreferential attachment mechanism [27], a network of retweets on Twitter [35] is explained by a regular preferential attachment mechanism, and the phylogenetic tree of West Nile virus [36] certainly did not grow by this mechanism. The credible intervals found when using only the static network overlap with the one we find when using available temporal metadata.