Confidence Regions and Averaging for Trees

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier,France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

Joint Work with
Louis Billera (Mathematics, Cornell)
Persi Diaconis (Mathematics and Statistics, Stanford)
Karen Vogtmann (Mathematics, Cornell).

Confidence Regions and Averaging for Trees

Combinatorics and Geometry for tree space

Susan Holmes
Statistics Department, Stanford
and INRA- Biométrie, Montpellier,France
susan@stat.stanford.edu
http://www-stat.stanford.edu/~susan/

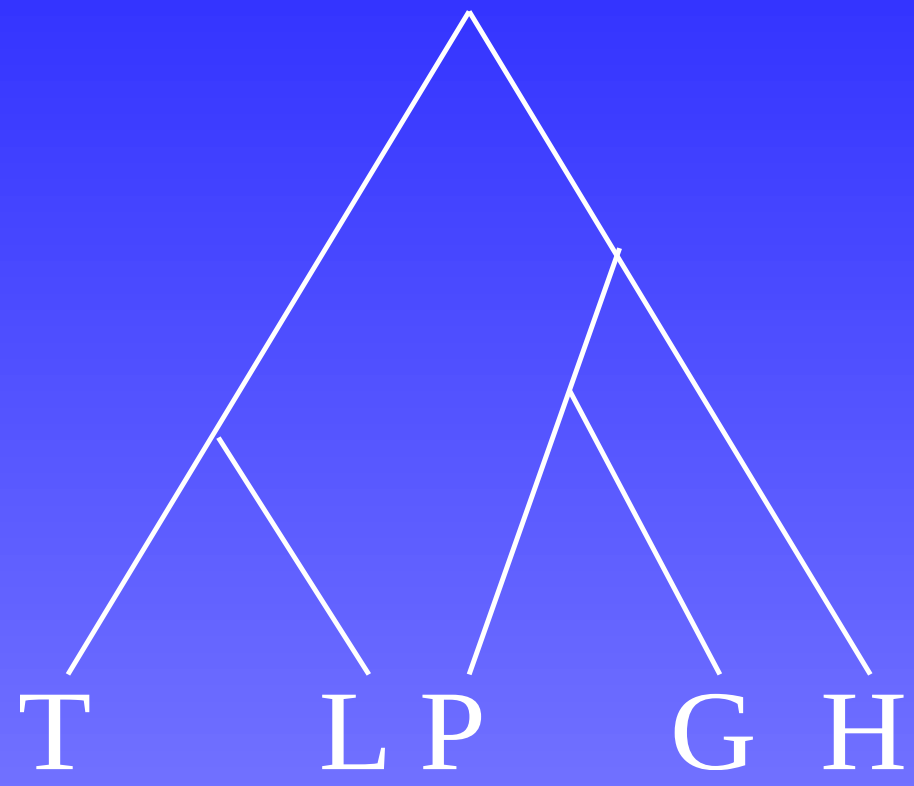
Joint Work with
Louis Billera (Mathematics, Cornell)
Persi Diaconis (Mathematics and Statistics, Stanford)
Karen Vogtmann (Mathematics, Cornell).

Tree for all

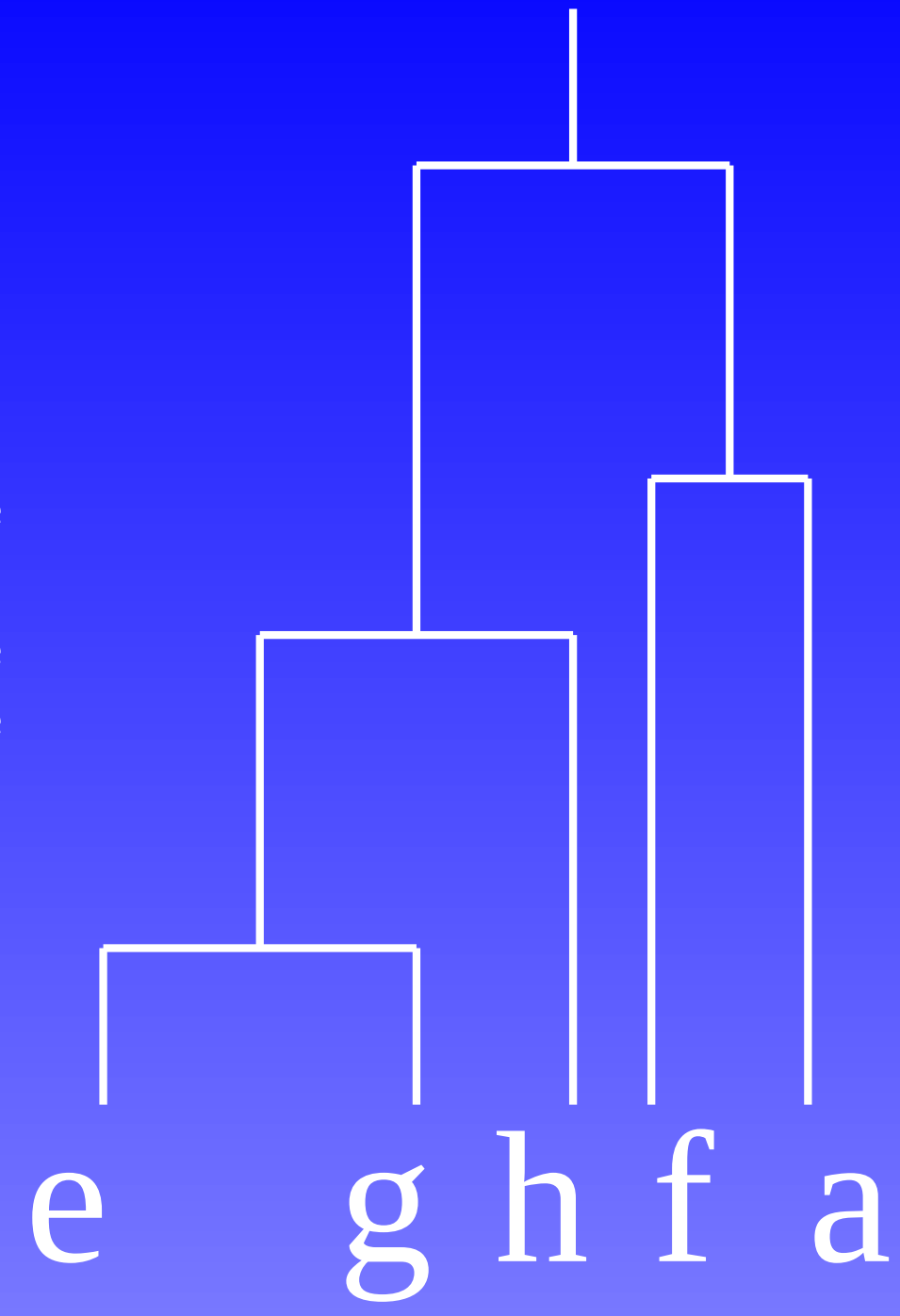
Tree for all

Where do collections of trees come from?

- Family trees or phylogenetic trees whose leaves are different evolutionary entities (species, genes, populations).

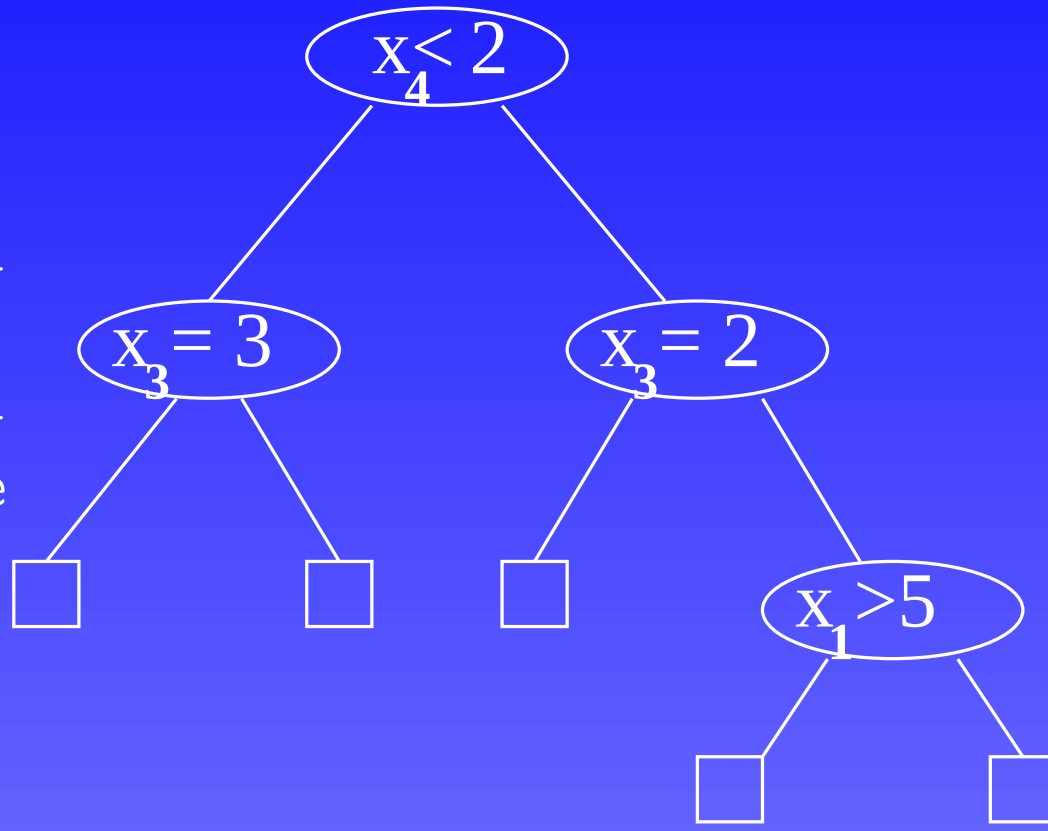


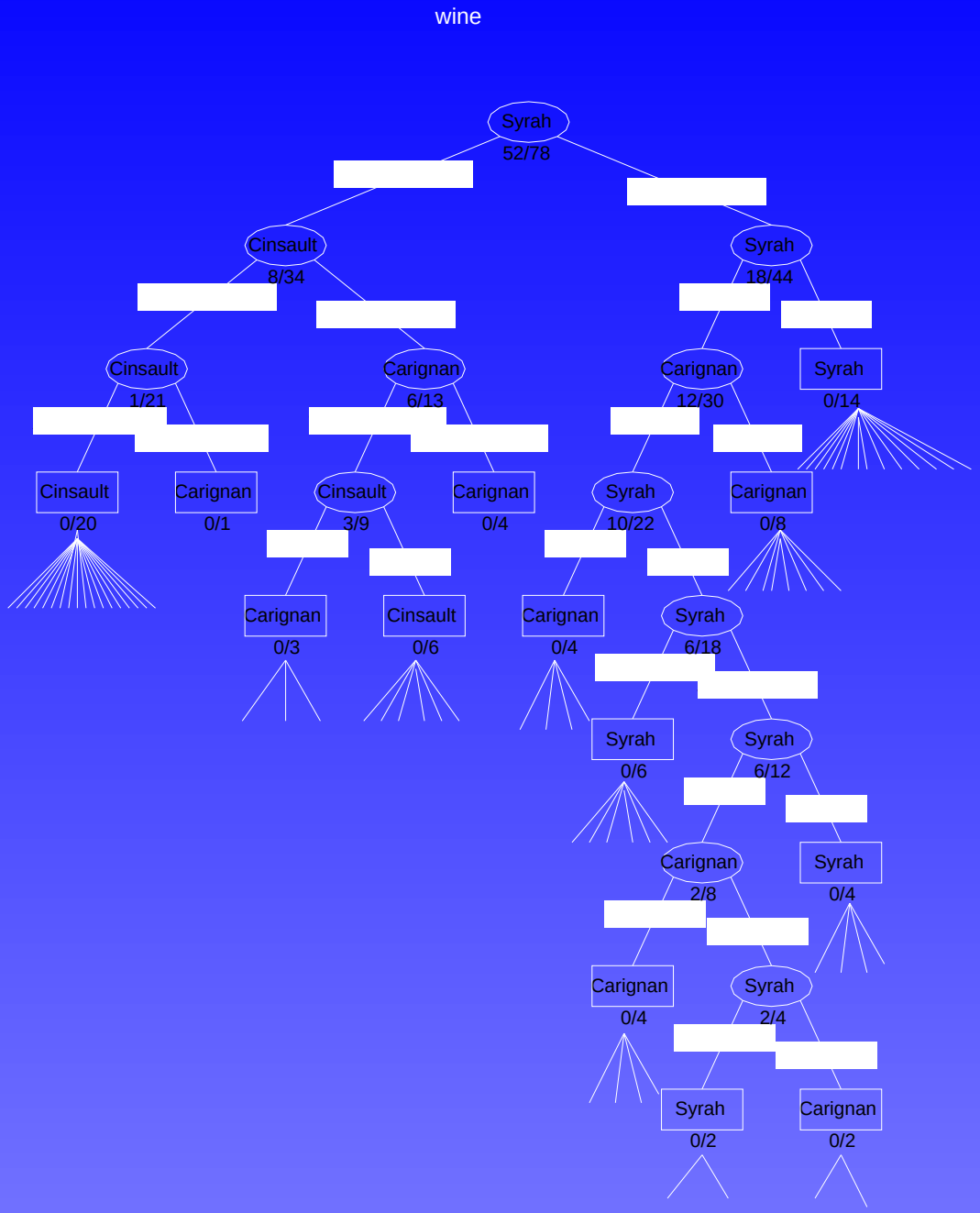
- Clustering trees for multiple clustering arrangements obtained over time for the same entities, cells or genes in gene expression for instance.



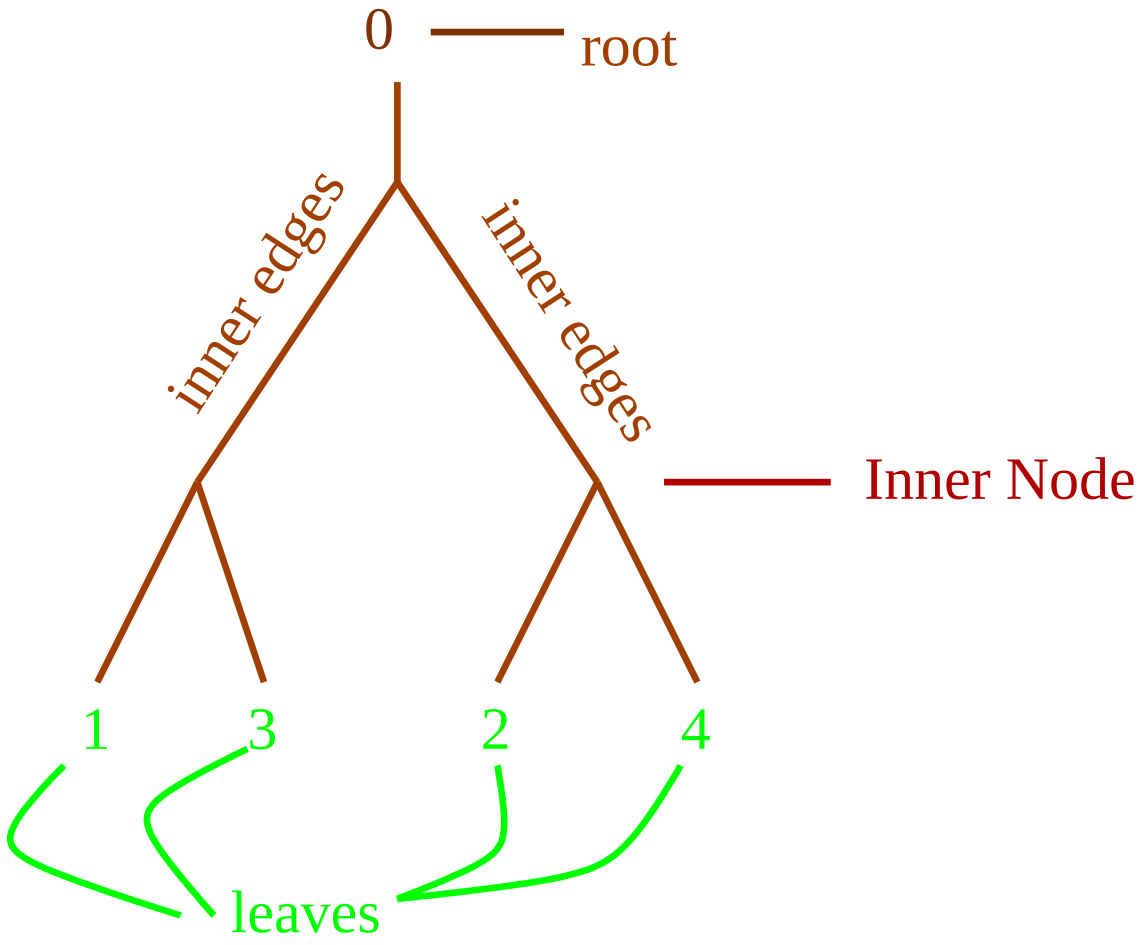
- Decision trees, such as Classification and Regression trees, whose leaves are the single observations before the trees have been pruned back.

- Decision trees, such as Classification and Regression trees, whose leaves are the single observations before the trees have been pruned back.

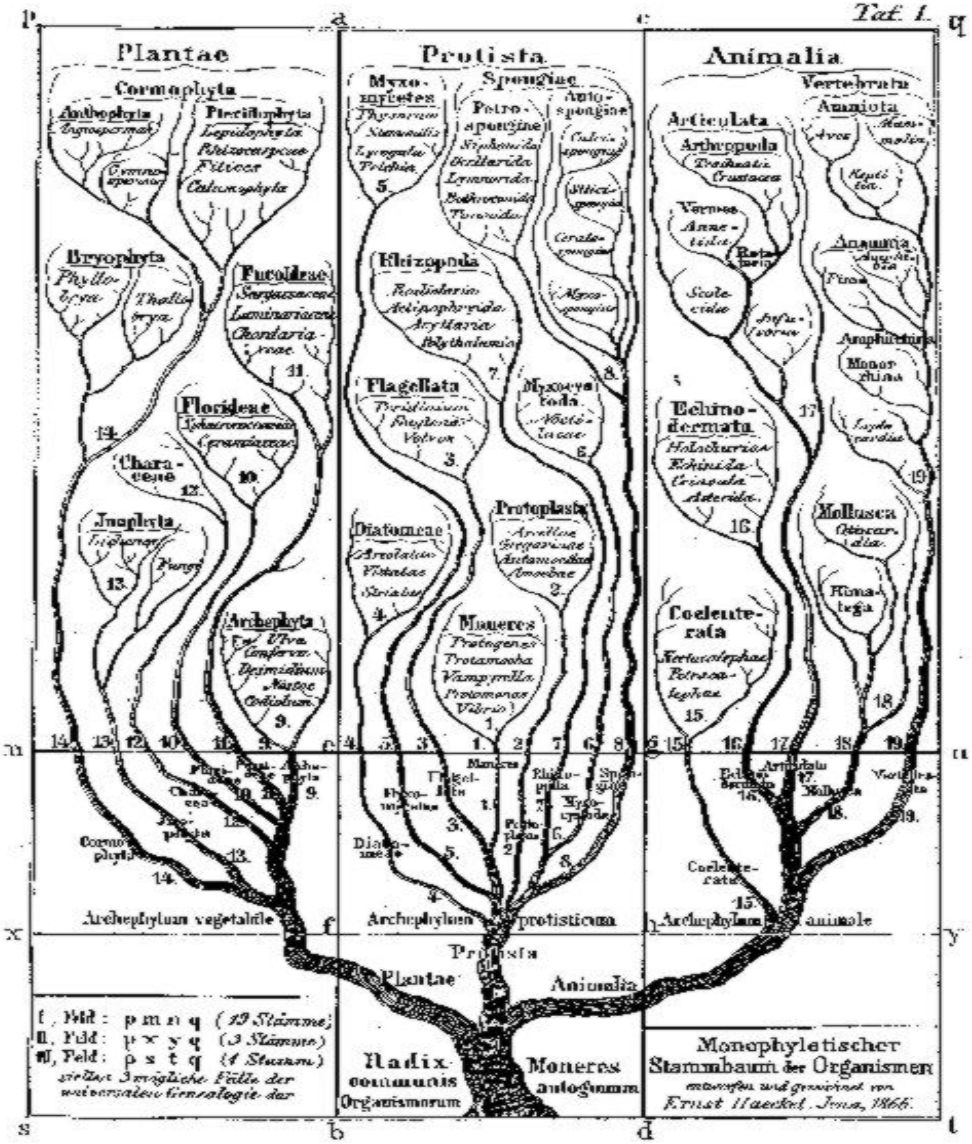



wine
Syrah
52/78
Cinsault
8/34
Syrah
18/44
Cinsault
1/21
Carignan
6/13
Carignan
12/30
Syrah
0/14
Cinsault
0/20
Carignan
0/1
Cinsault
3/9
Carignan
0/4
Syrah
10/22
Carignan
0/8
Carignan
0/3
Cinsault
0/6
Carignan
0/4
Syrah
6/18
Syrah
0/6
Syrah
6/12
Carignan
2/8
Syrah
0/4
Carignan
0/4
Syrah
2/4
Syrah
0/2
Carignan
0/2

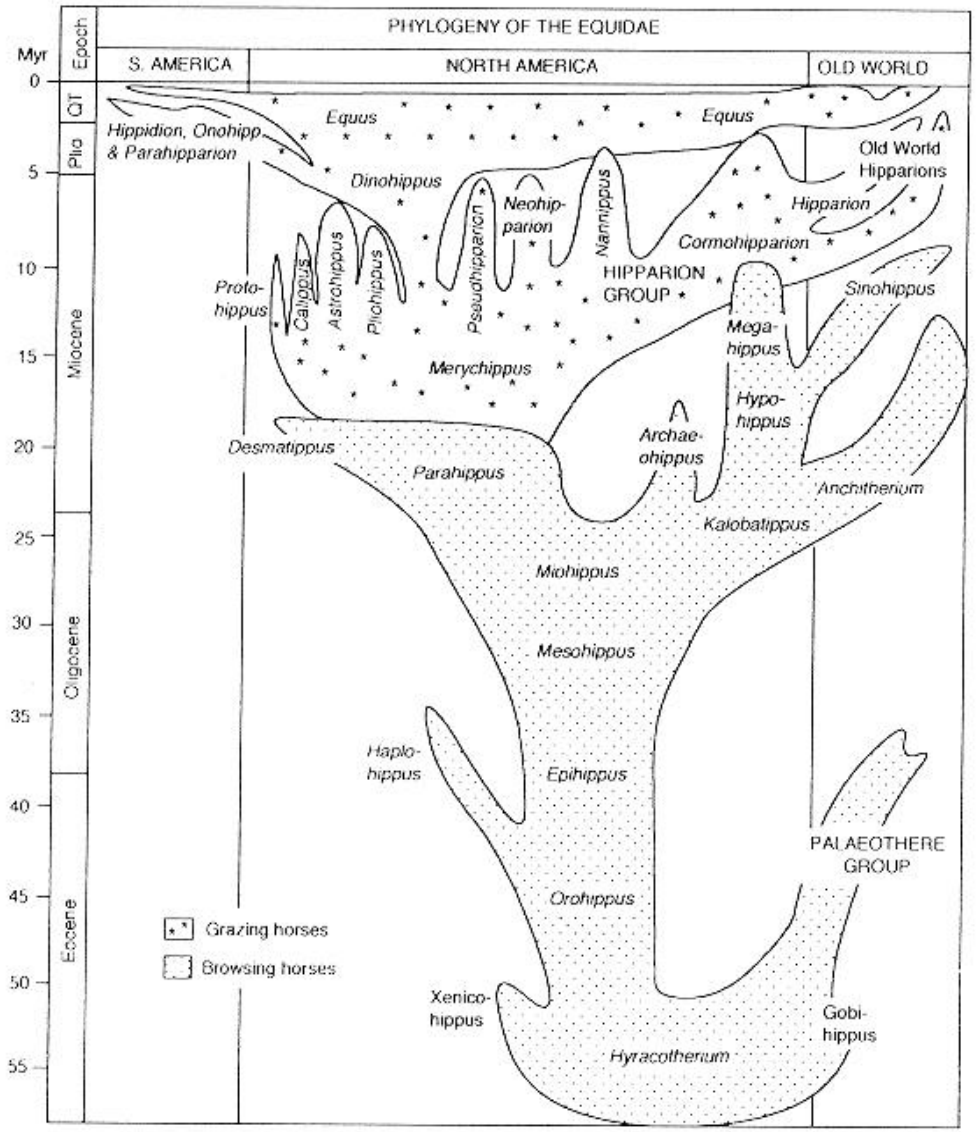
A semi-labeled binary Tree



Phylogenetic Trees



Confidence Statements for trees



DNA Data for 12 species of primates

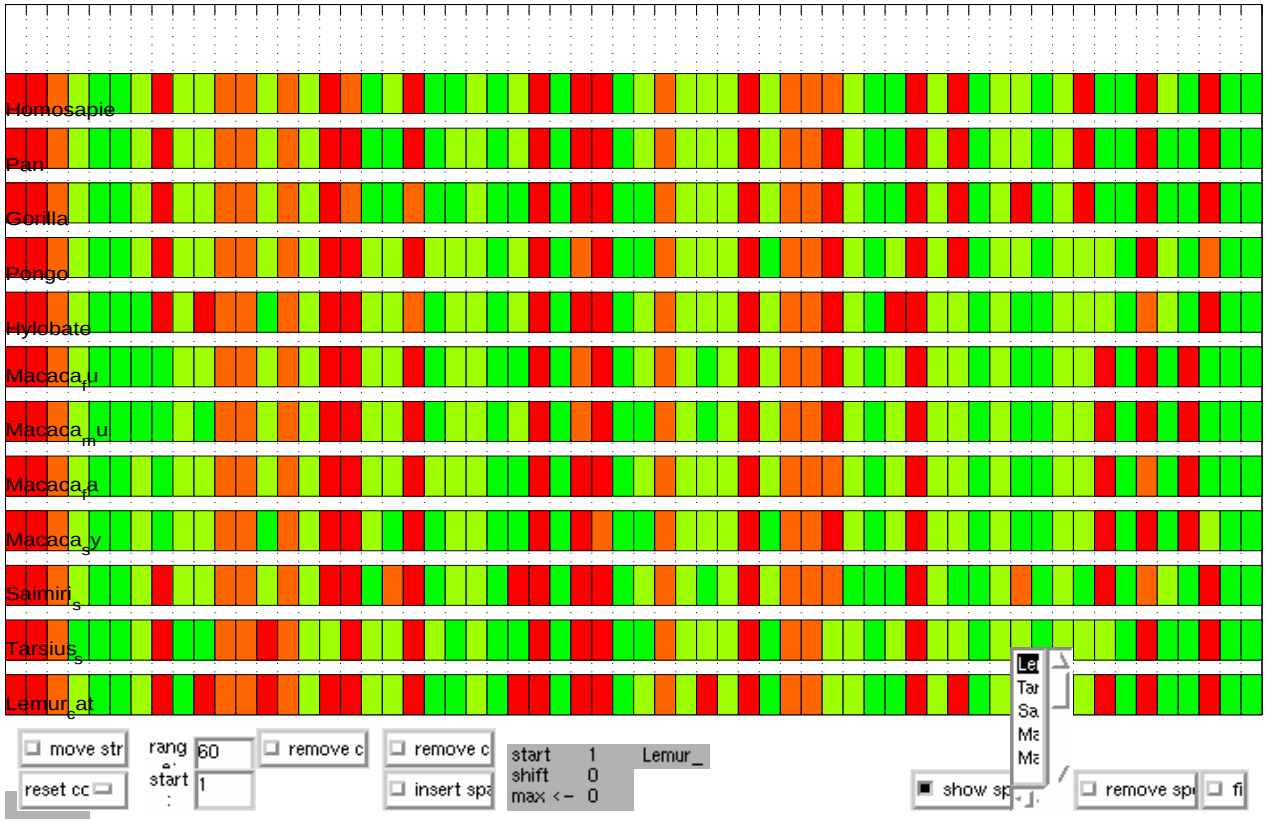
Mitochondria, 898 characters on 12 species.

Hayasaka, K., T. Gojobori, and S. Horai. 1988.

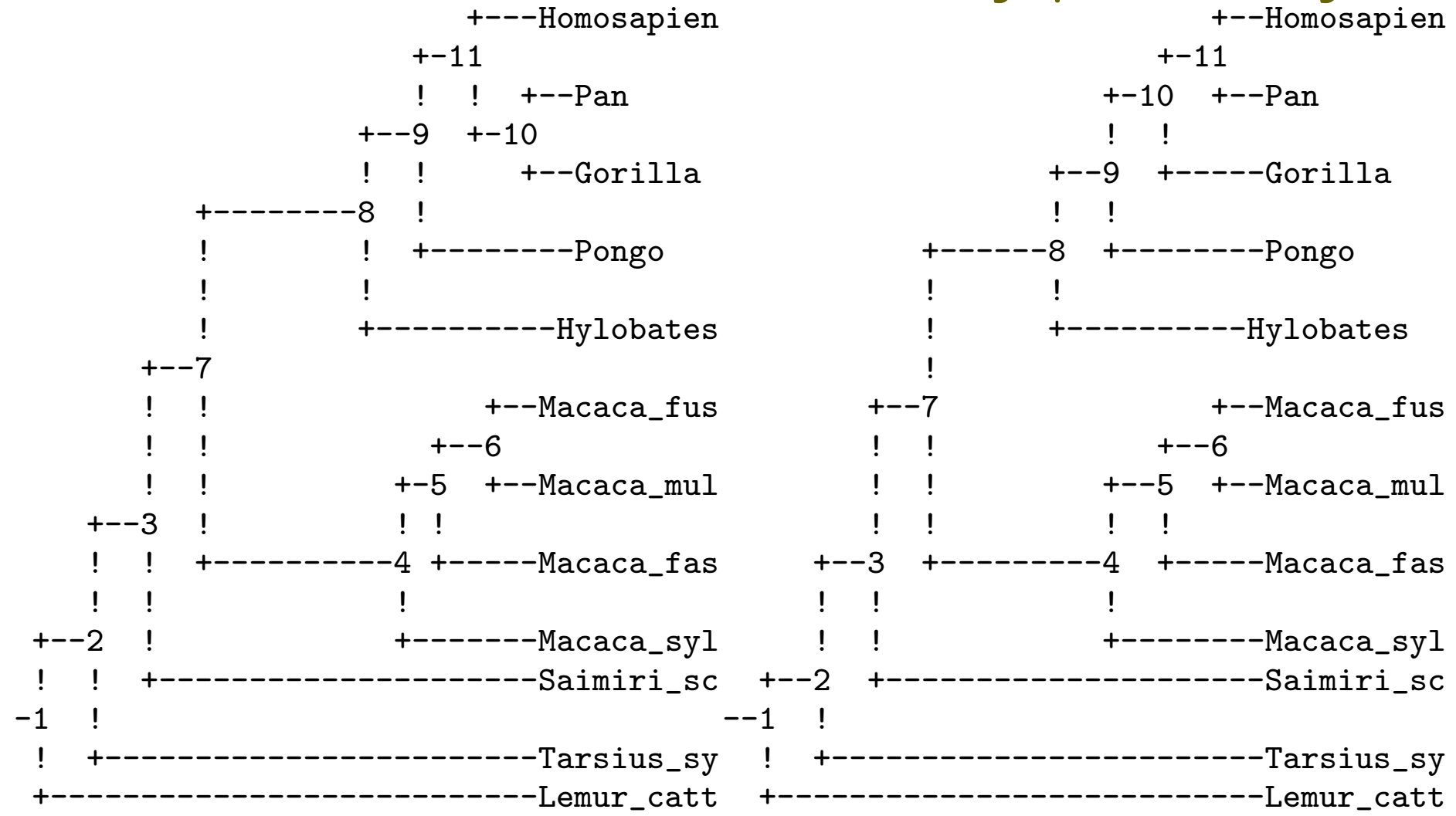
Trees are built from DNA data such as the following:

```
  12  60
Lemur_cat AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC CTTACATCAT CCATATT
Tarsius_s AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC CTCACCTCCT CCCTATT
Saimiri_s AAGCTTCACC GGCGCAATGA TCCTAATAAT CGCTCACGGG TTTACTTCGT CTATGCT
Macaca_sy AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA CTCACCTCTT CCATATA
Macaca_fa AAGCTTCTCC GGCGCAACCA CCCTTATAAT CGCCCACGGG CTCACCTCTT CCATGTA
Macaca_mu AAGCTTTTCT GGCGCAACCA TCCTCATGAT TGCTCACGGA CTCACCTCTT CCATATA
Macaca_fu AAGCTTTTCC GGCGCAACCA TCCTTATGAT CGCTCACGGA CTCACCTCTT CCATATA
 Hylobate AAGCTTTACA GGTGCAACCG TCCTCATAAT CGCCCACGGA CTAACCTCTT CCCTGCT
Pongo     AAGCTTCACC GGCGCAACCA CCCTCATGAT TGCCCATGGA CTCACATCCT CCCTACT
Gorilla   AAGCTTCACC GGCGCAGTTG TTCTTATAAT TGCCCACGGA CTTACATCAT CATTATT
Pan       AAGCTTCACC GGCGCAATTA TCCTCATAAT CGCCCACGGA CTTACATCCT CATTATT
Homosapie AAGCTTCACC GGCGCAGTCA TTCTCATAAT CGCCCACGGG CTTACATCCT CATTACT
```

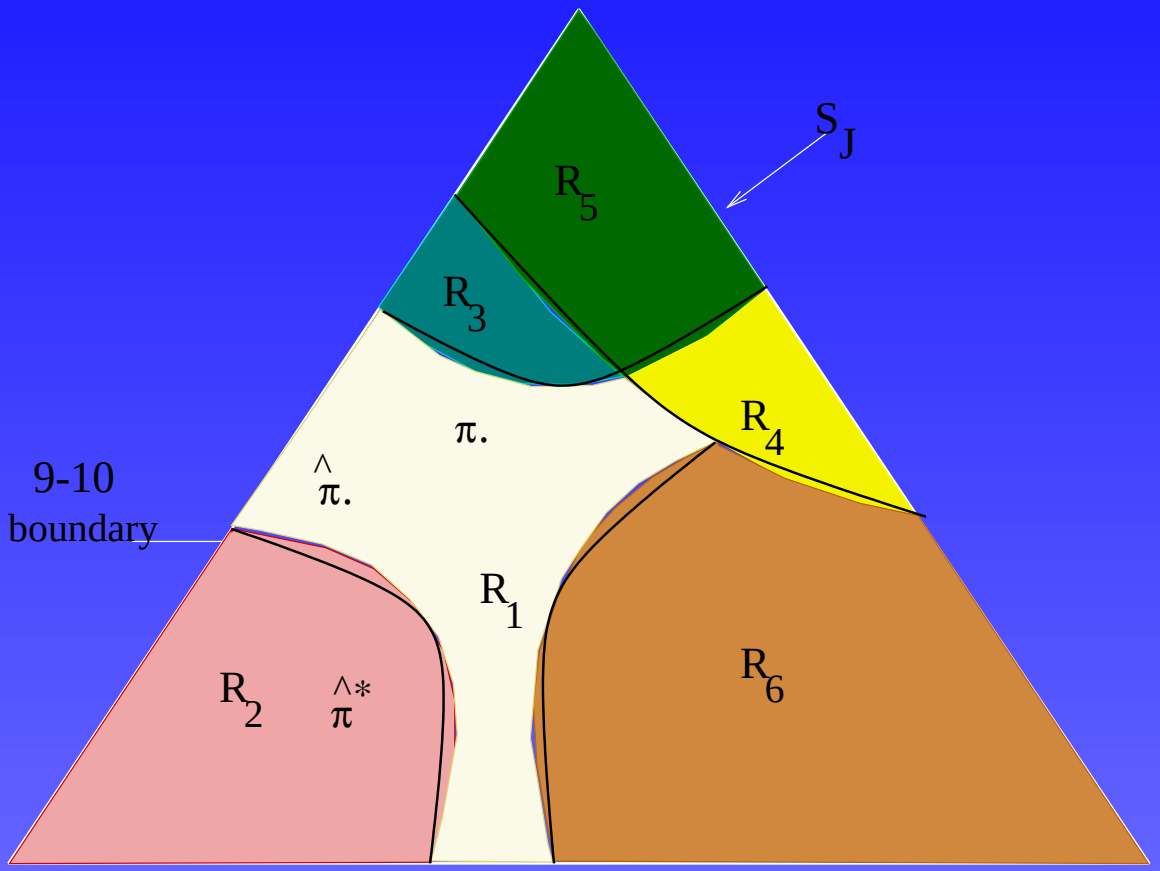
Color Coded version of the data, after alignment



Two Trees Built with this data by parsimony

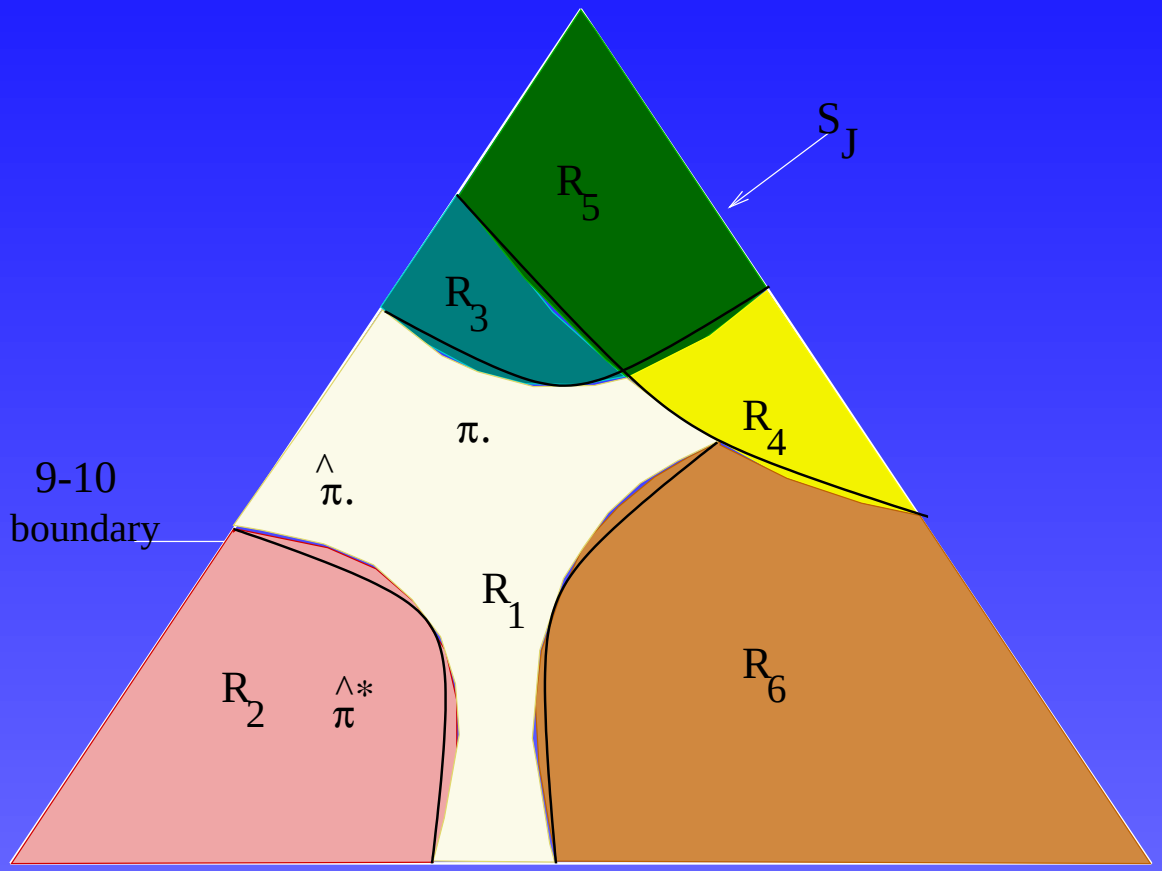


Confidence Regions



From Efron, Halloran, Holmes, (1996)[0].

Confidence Regions



From Efron, Halloran, Holmes, (1996)[0].

What is the curvature of the boundary?

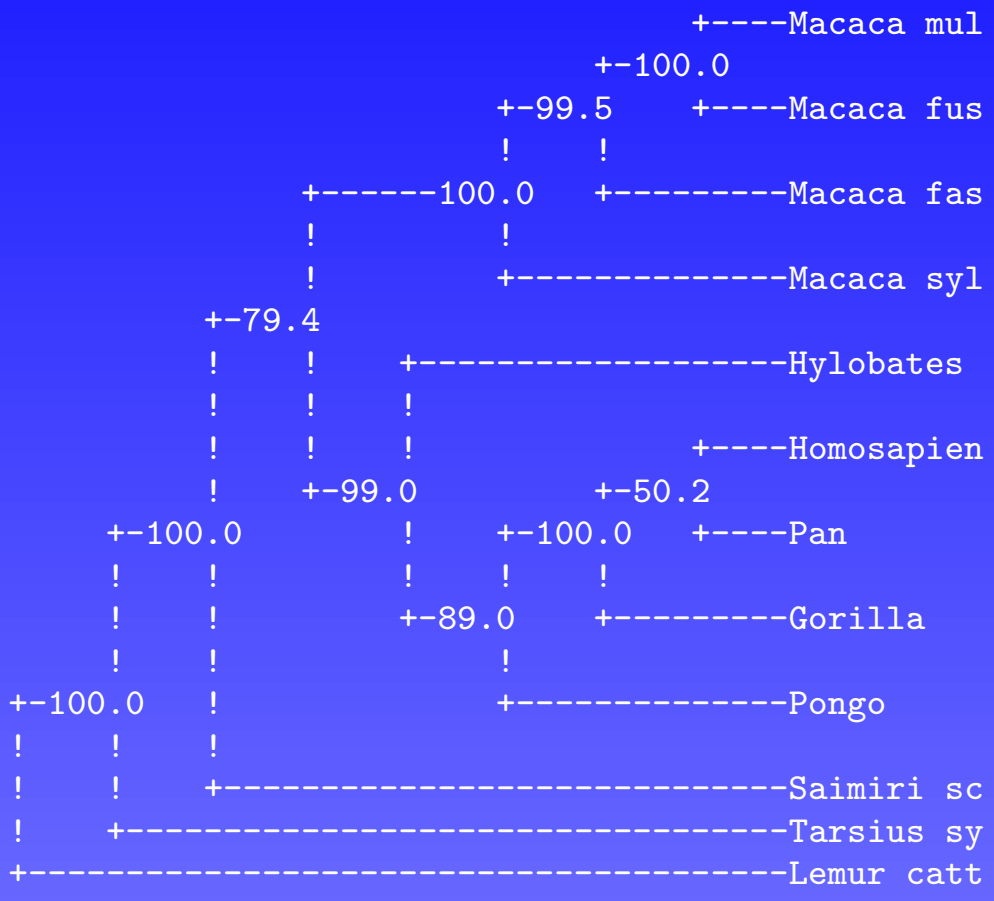
How many neighbors does a region have?

Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.

Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.



Classification And Regression Trees

Bootstrap for CART trees.
Average Tree seems to be much better than one tree built with all the data. (Amit and Geman (1998))

What are the ways of averaging trees?
Consensus majority rule or strict consensus.

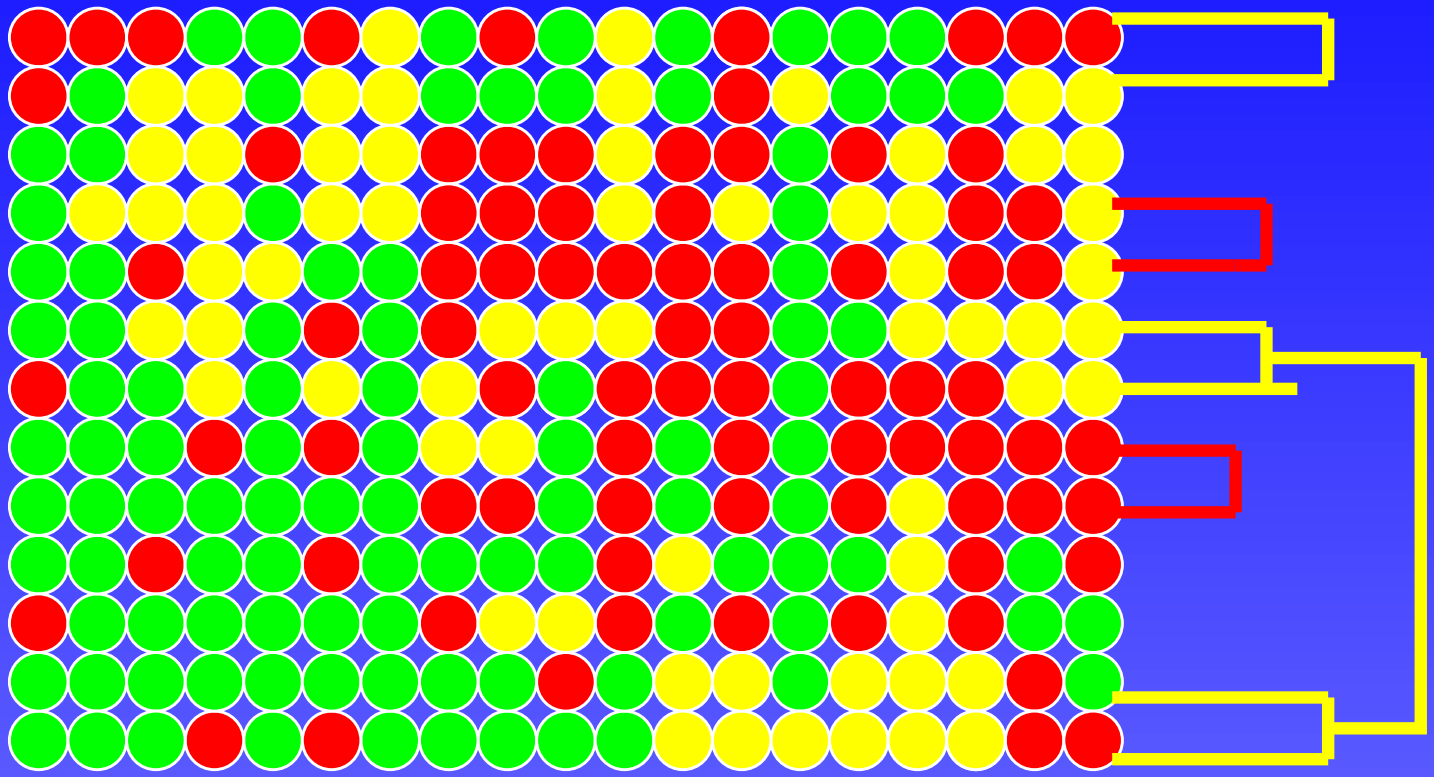
Clustering Trees

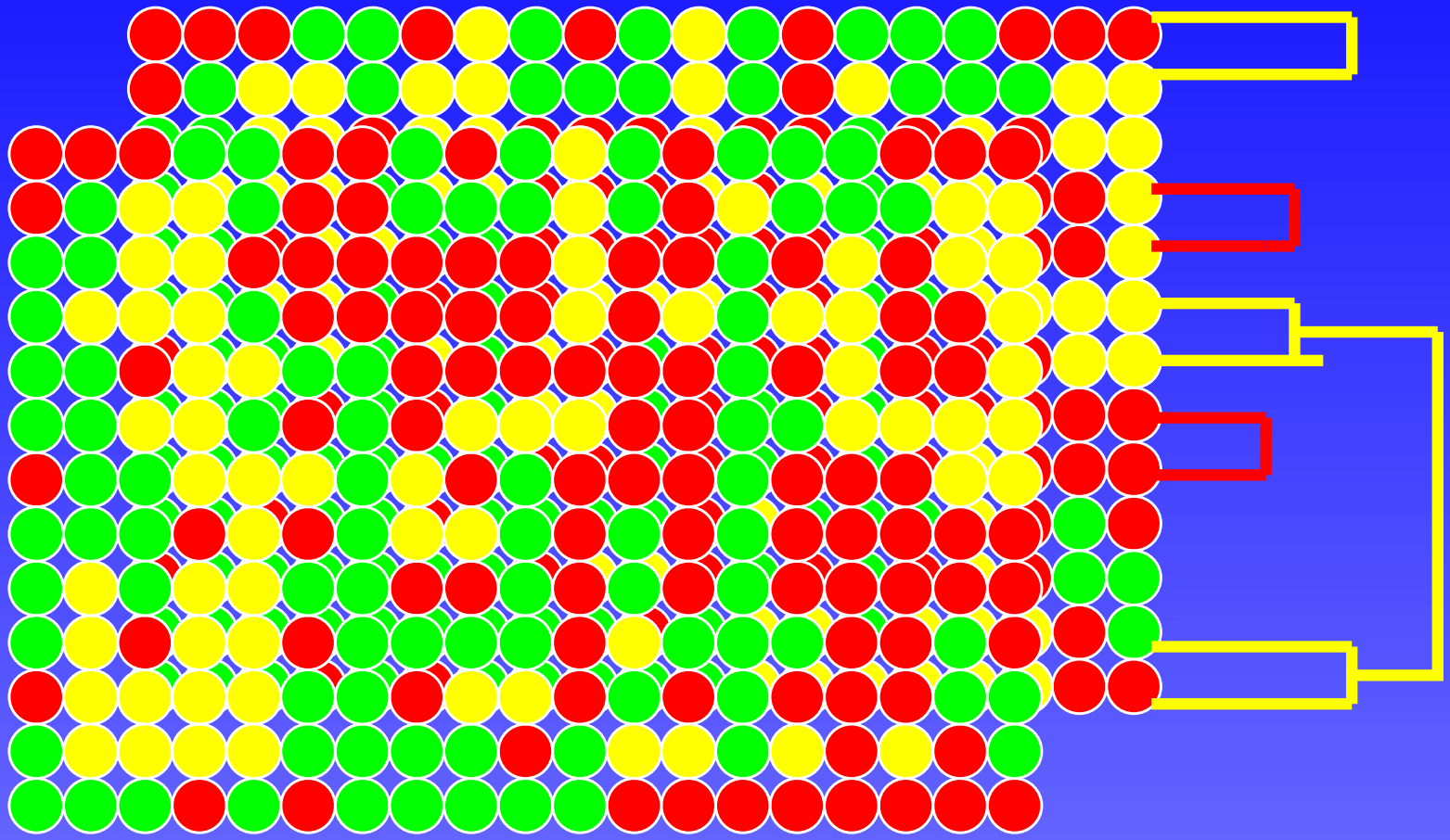
One hierarchical cluster per matrix.

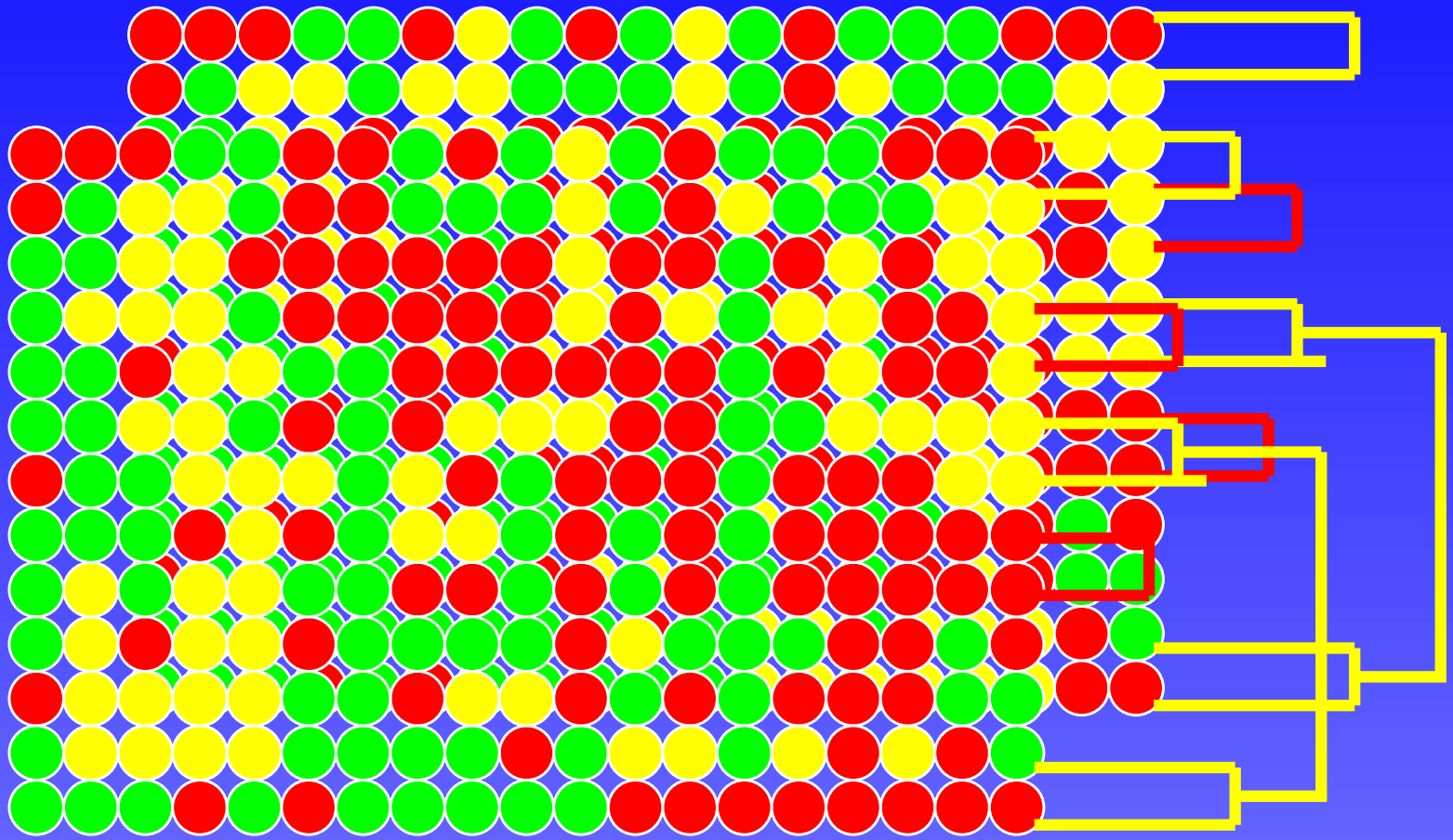
Cube of matrices.

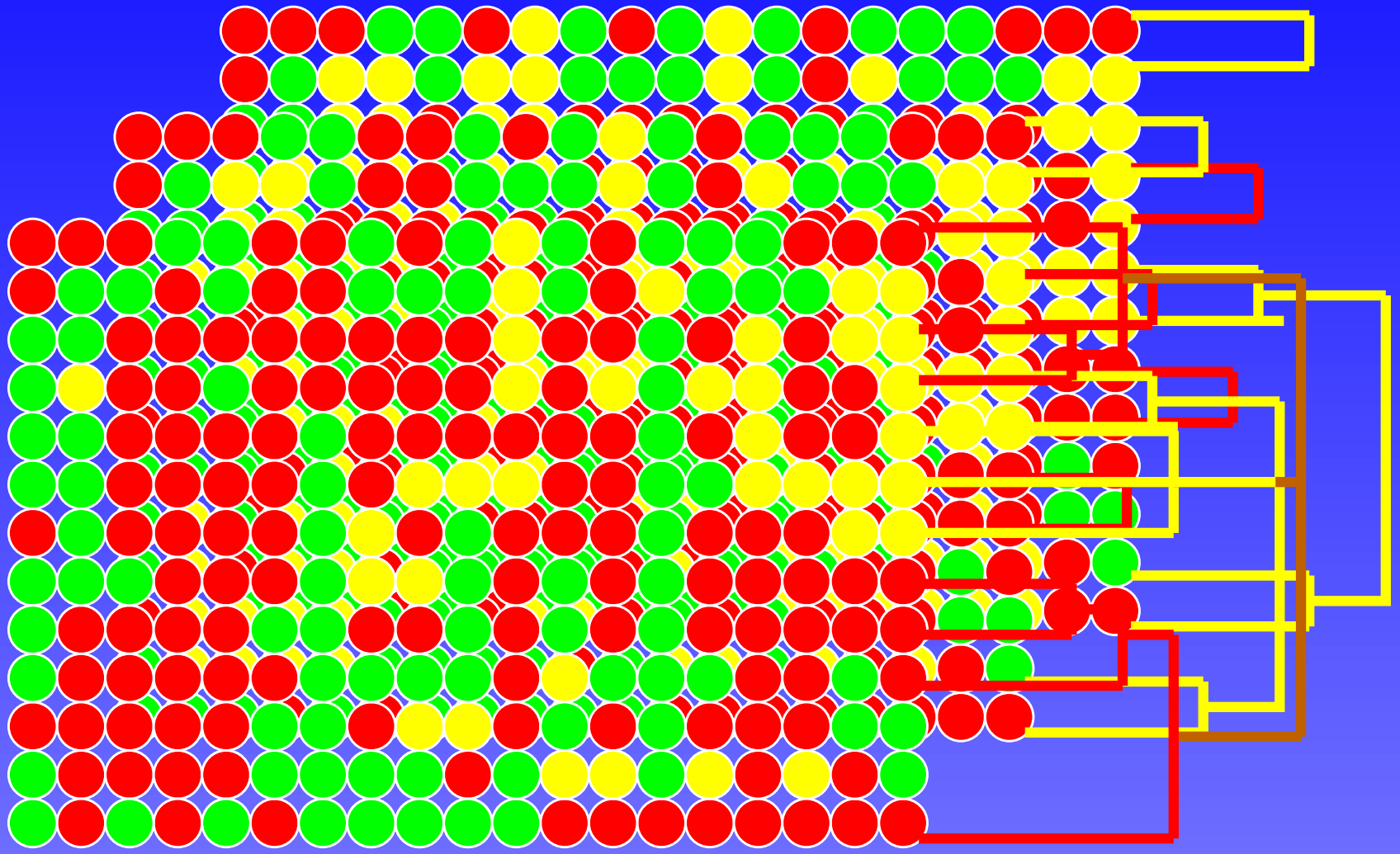
How do we flatten or follow in time the trees as they ‘change’, we want a space for following the pathways of trees.

Clustering Trees
For micro-arrays (Eisen et al, 1999)









Binary Trees

Other parameterizations for Tree Space

Doss, Pearl, Li , *JASA to appear*(2000) [0]. Shannon and Banks , *Statistics in Medecine* (1997) [0] Mau, Larget and Newton, *Biometrika* (1998) [0].

We would like to fill in the matching polytope.

Matchings and Binary Semi-labeled Trees

This comes from Diaconis and Holmes (1998) A matching of 2(n-1) objects is a pairing off, without care for order within pairs or between pairs.

The Same matchings:

$$(1,4)(2,5)(3,6)$$
$$(6,3)(4,1)(2,5)$$
$$(5,2)(3,6)(1,4)$$

Matchings and Binary Semi-labeled Trees

This comes from Diaconis and Holmes (1998) A matching of 2(n-1) objects is a pairing off, without care for order within pairs or between pairs.

The Same matchings:

$$(1,4)(2,5)(3,6)$$
$$(6,3)(4,1)(2,5)$$
$$(5,2)(3,6)(1,4)$$

Call $\mathcal{B}_{n-1}$ the subgroup of $\mathcal{S}_{2n-2}$ that fixes the pairs

$$\{1,2\}\{3,4\}\ldots\{2n-3,2n-2\}$$

then

$$\mathcal{M}_{n-1} = \mathcal{S}_{2n}/\mathcal{B}_{n-1}$$

and

$$|\mathcal{M}_{n-1}| = \frac{(2n-2)!}{2^{n-1}(n-1)!} = (2n-3)!! = (2n-3)\times(2n-5)\times\cdots\times 3\times 1$$

This formula for the number of trees was first proved using generating functions by Schroder (1873)[0].

$(\mathcal{S}_{2n-2}, \mathcal{B}_{n-1})$ form a Gelfand pair Diaconis and Shahshahani (1987) [**?**].

$$L(\mathcal{M}_{n-1}) = V_1 \oplus V_2 \oplus \ldots \oplus V_\lambda$$

A multiplicity free representation.

$$L(\mathcal{M}_{n-1}) = \bigoplus_{\lambda \vdash n} \mathcal{S}^{2\lambda}$$

where the direct sum is over all partitions λ of m, $2\lambda = (2\lambda_1, 2\lambda_2, \ldots, 2\lambda_k)$ and $\mathcal{S}^{2\lambda}$ is associated irreducible representation of the symmetric group S_{2m}.

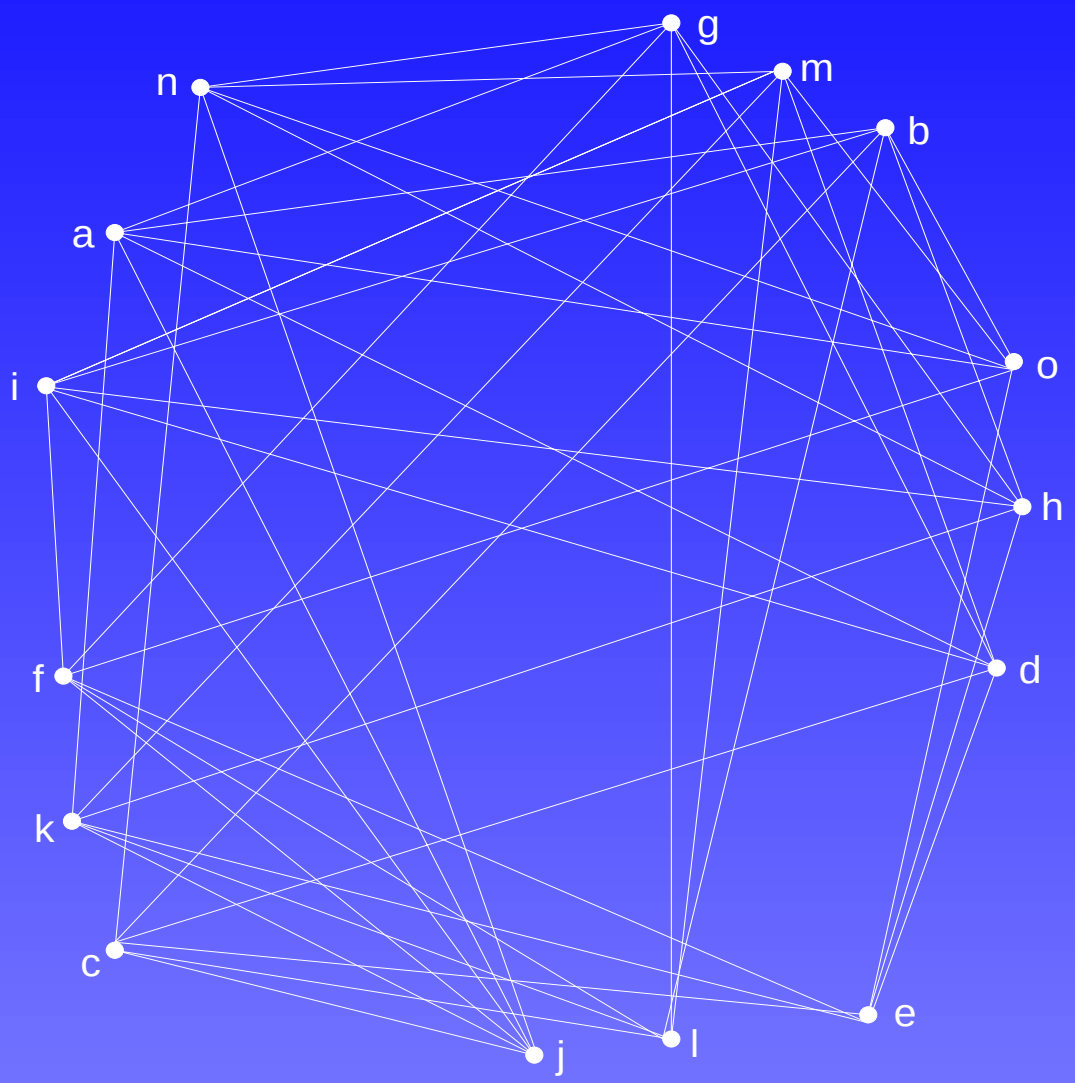
Just to take the first few: for $\lambda = n-1$ S^λ are the constants, and this gives the sample size. for $\lambda = (n-2, 1)$, S^λ are the number of times each pair appears. for $\lambda = (n-3, 2)$, S^λ are the number of times partition of 4 appears in the tree. for $\lambda = (n-3, 1, 1)$, S^λ are the number of times 2 pairs appear simultaneously. This decomposition is similar to what was done by Diaconis for permutation data.[0]

Matchings are useful

- For going through all trees systematically. (Gray code [0] for Trees)
- Doing vigorous random walks on tree space.
- Doing Fourier Analysis on Tree Data.

But the matching distance is not satisfactory to the biologists.

The Matching Polytope



Mallow's Model

$P(\tau) \ =$

Mallow's Model

$P(\tau) = Ke^{-\lambda d(\tau,\tau_0)}$, K is a normalizing constant

Mallow's Model

$$P(\tau) \;=\; Ke^{-\lambda d(\tau,\tau_0)}, \qquad K \text{ is a normalizing constant}$$

- Exponential family, it needs:

Mallow's Model

$$P(\tau) \;=\; Ke^{-\lambda d(\tau,\tau_0)}, \qquad K \text{ is a normalizing constant}$$

- Exponential family, it needs:
  - A central tree τ_0
  - A distance between trees $d(\tau, \tau_0)$

Mallow's Model

$$P(\tau) \;=\; Ke^{-\lambda d(\tau,\tau_0)}, \qquad K \text{ is a normalizing constant}$$

- Exponential family, it needs:
  - A central tree τ_0
  - A distance between trees $d(\tau, \tau_0)$
- It is possible to extend this to make a Bayesian model with a symmetric apriori distribution for τ_0.

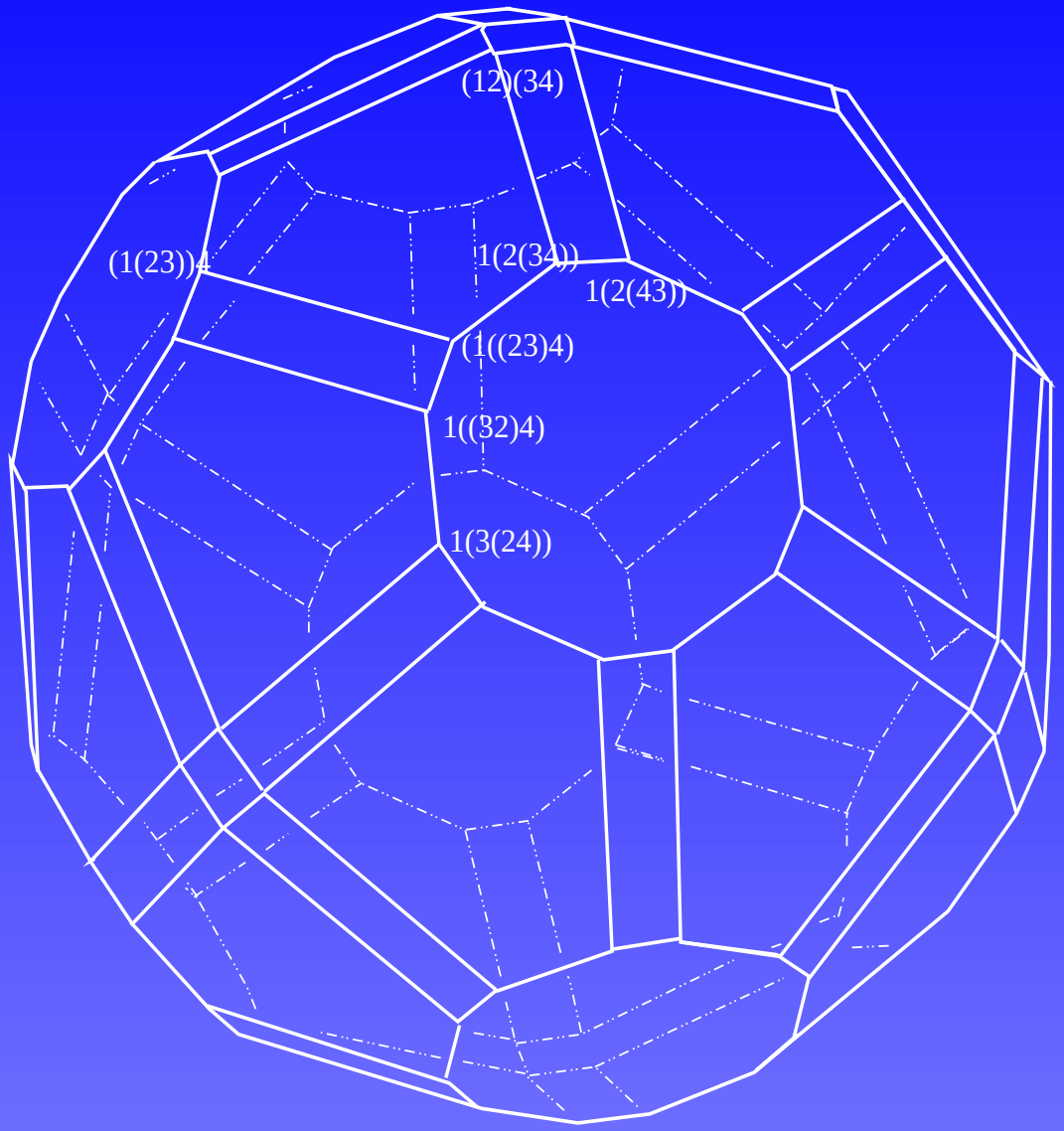
Mallow's Model

$$P(\tau) \;=\; Ke^{-\lambda d(\tau,\tau_0)}, \qquad K \text{ is a normalizing constant}$$

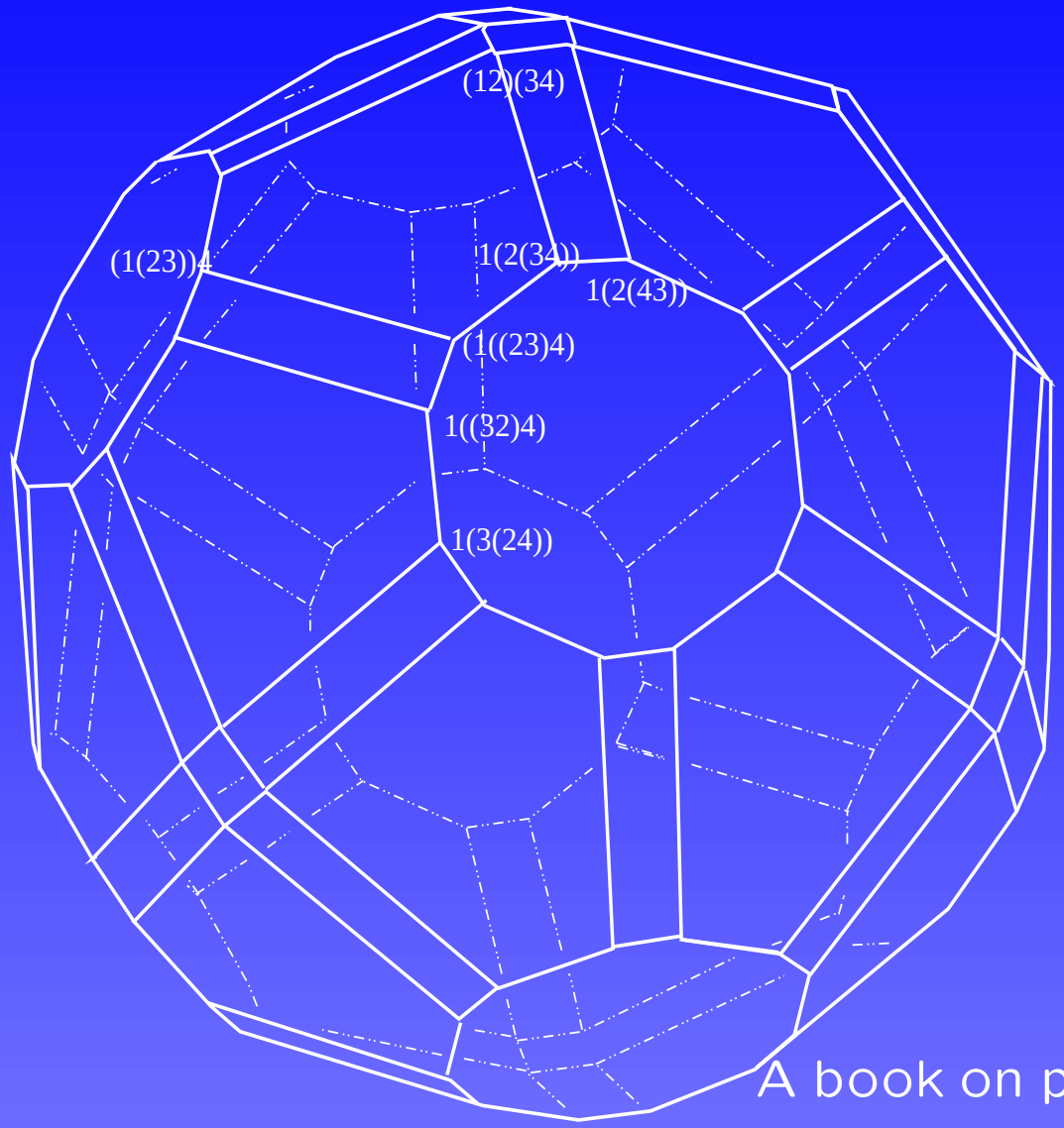
- Exponential family, it needs:
  - A central tree τ_0
  - A distance between trees $d(\tau, \tau_0)$
- It is possible to extend this to make a Bayesian model with a symmetric apriori distribution for τ_0.

Distances and centroids are essential

The permuto-associahedron



The permuto-associahedron



A book on polytopes.(Ziegler)

[0]

But the trees are extreme points

- A neigbor relation must be biologically acceptable (strong symmetry requirement)

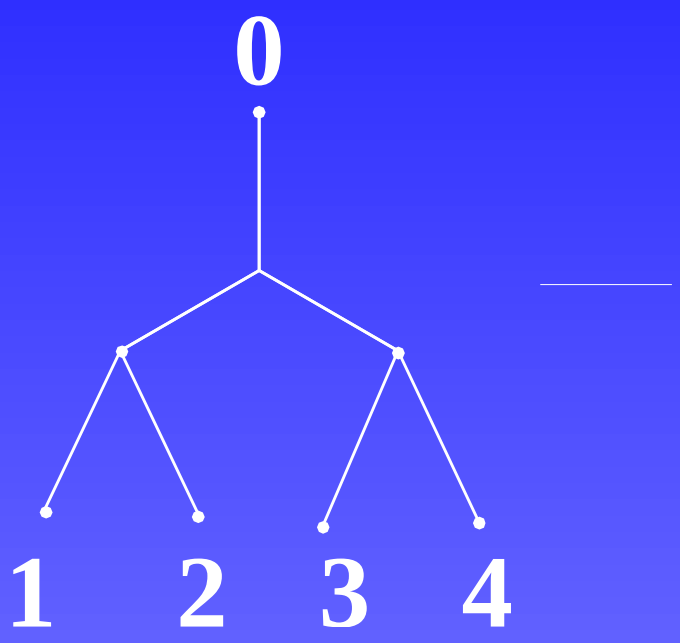
- The trees must sit in the center of the regions.

- A neigbor relation must be biologically acceptable (strong symmetry requirement)

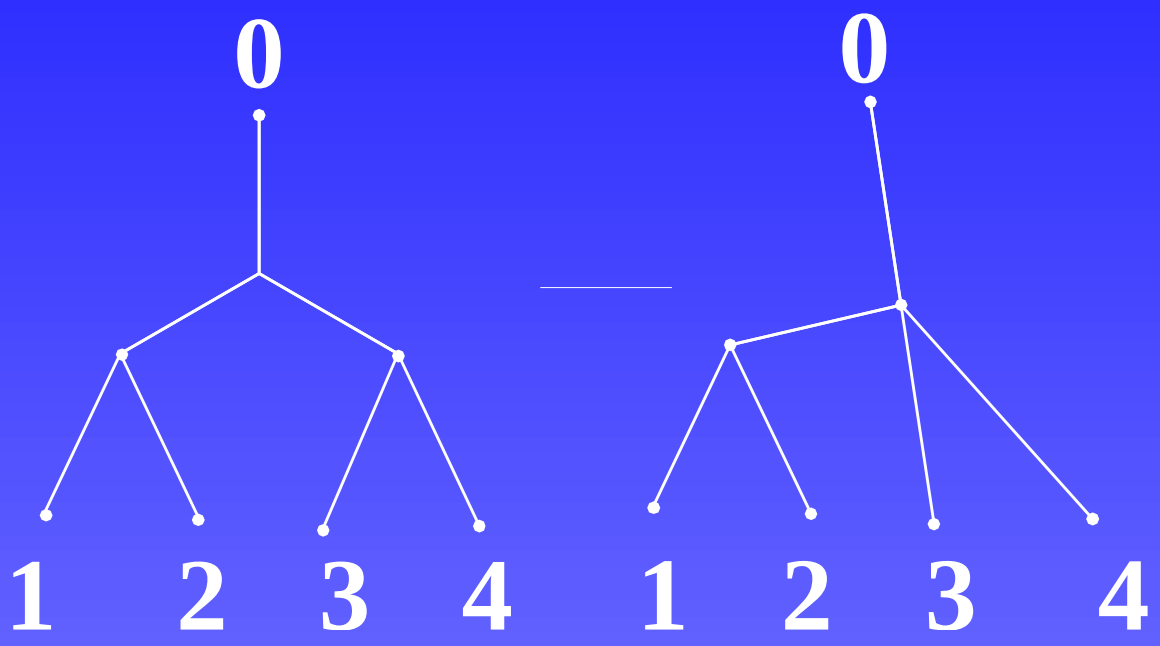
- The trees must sit in the center of the regions.

We could take the quotient of the polytope, but a direct construction is easier to visualize.

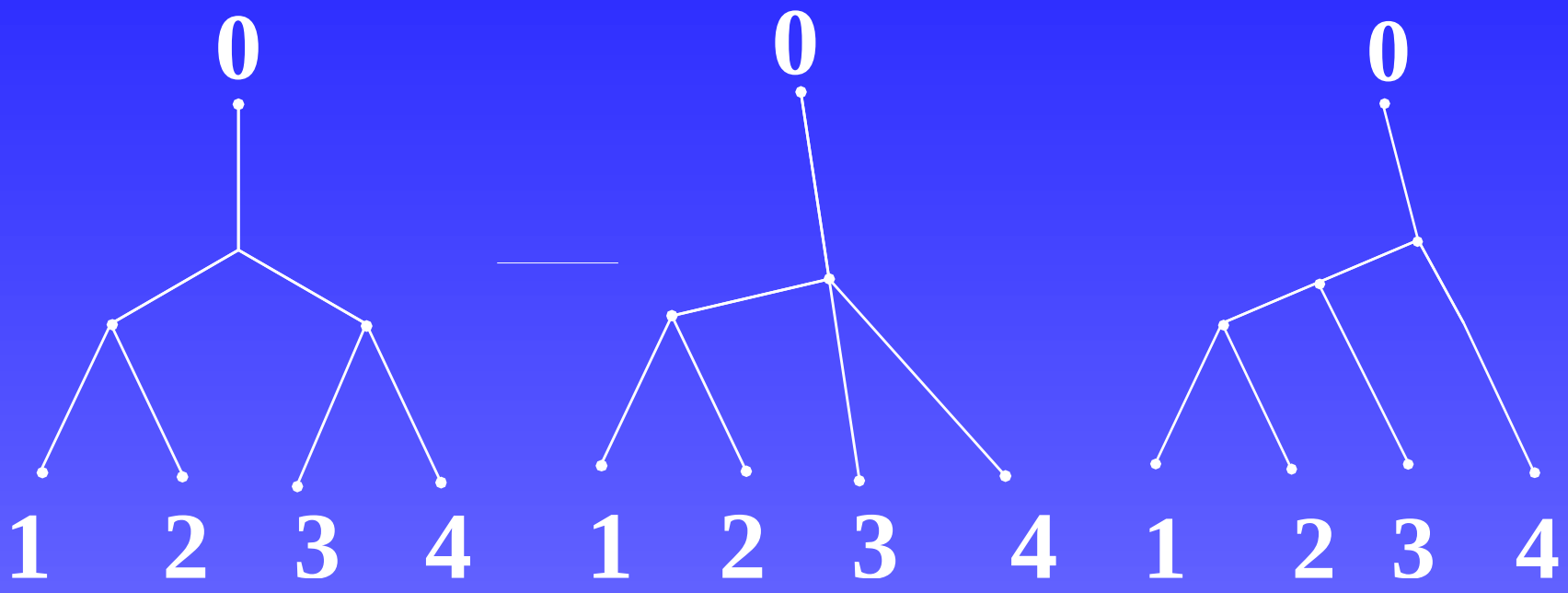
Rotation Moves



Rotation Moves

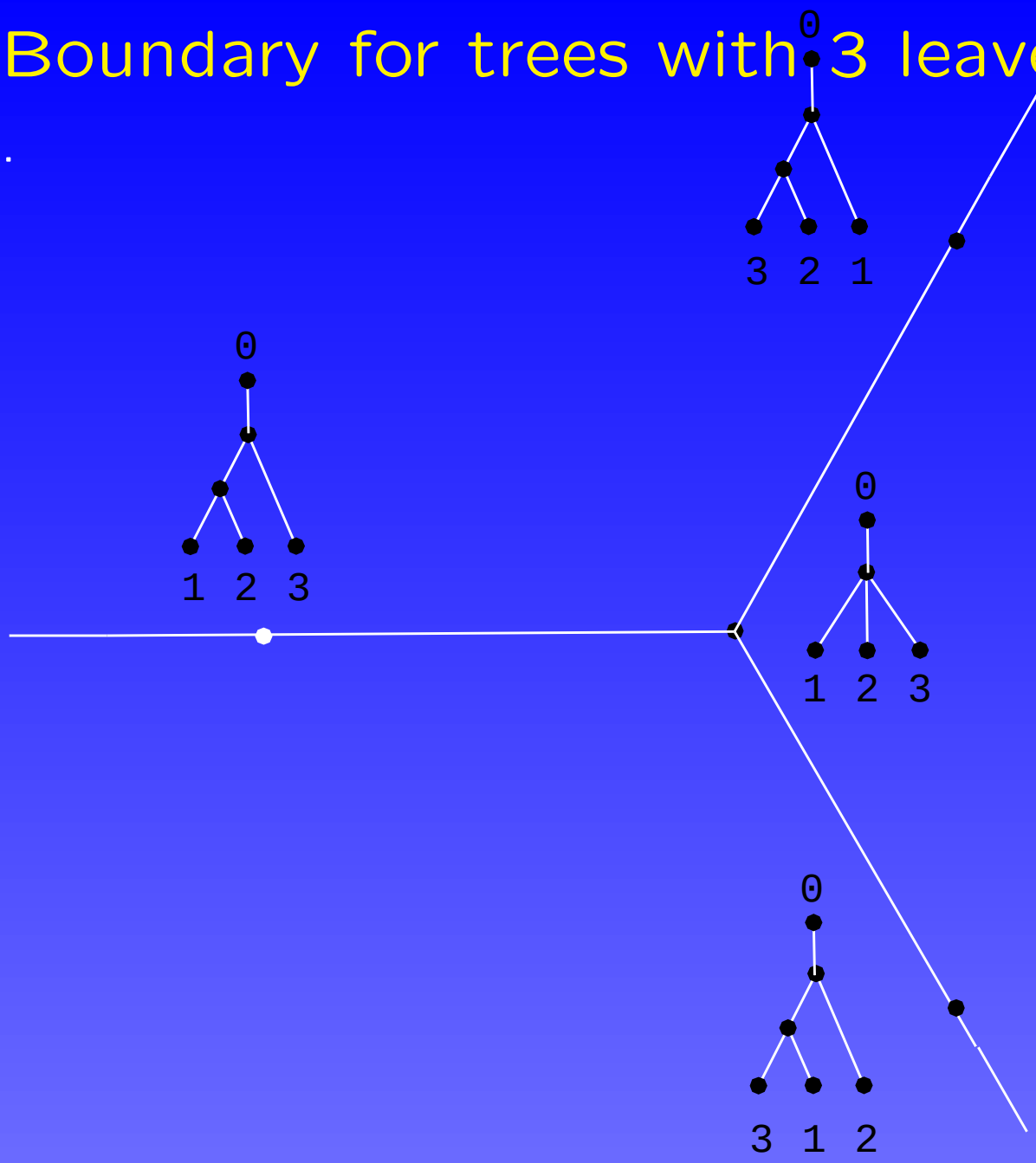


Rotation Moves

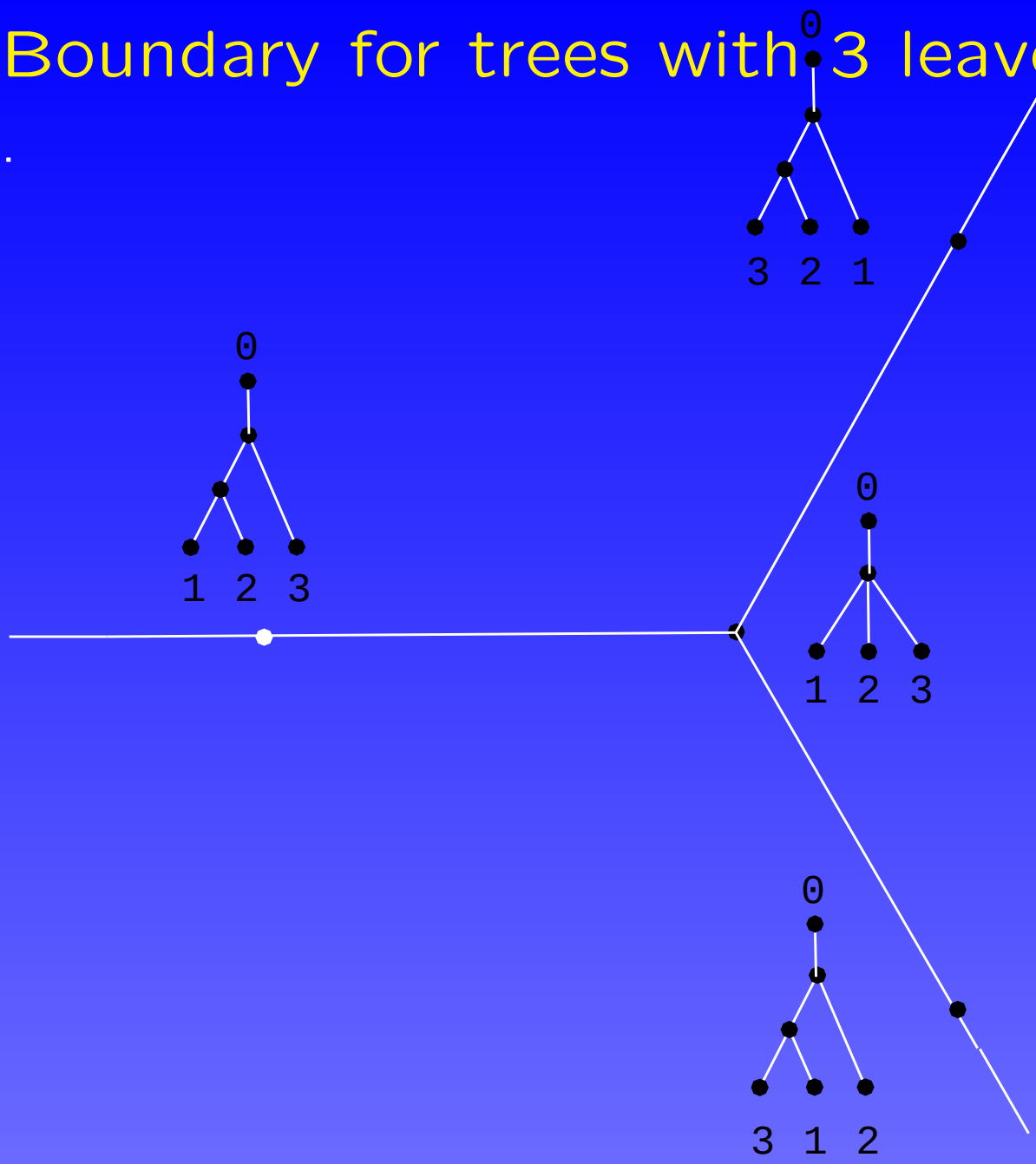


This is known to the biologists as the NNI moves.

Boundary for trees with 3 leaves

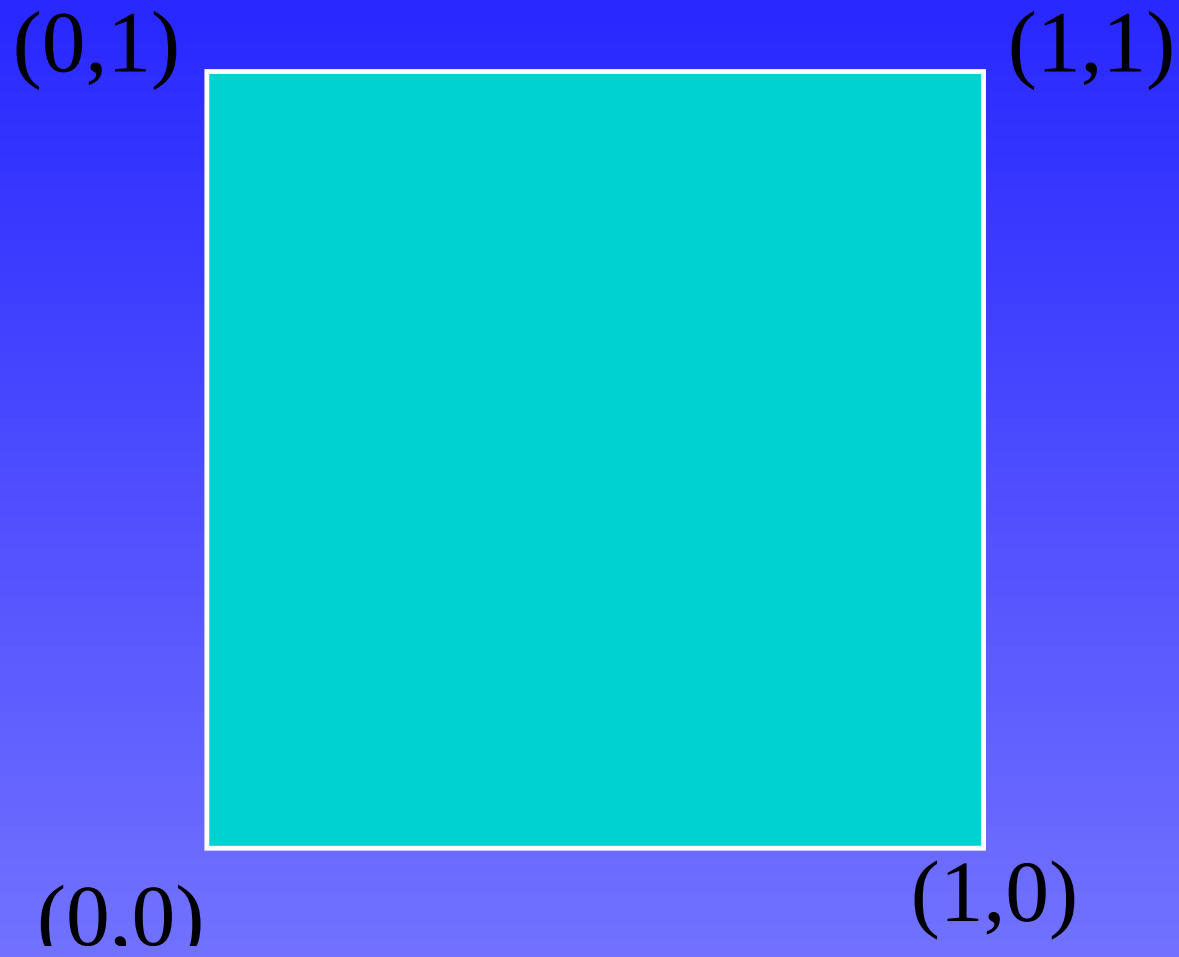


Boundary for trees with 3 leaves

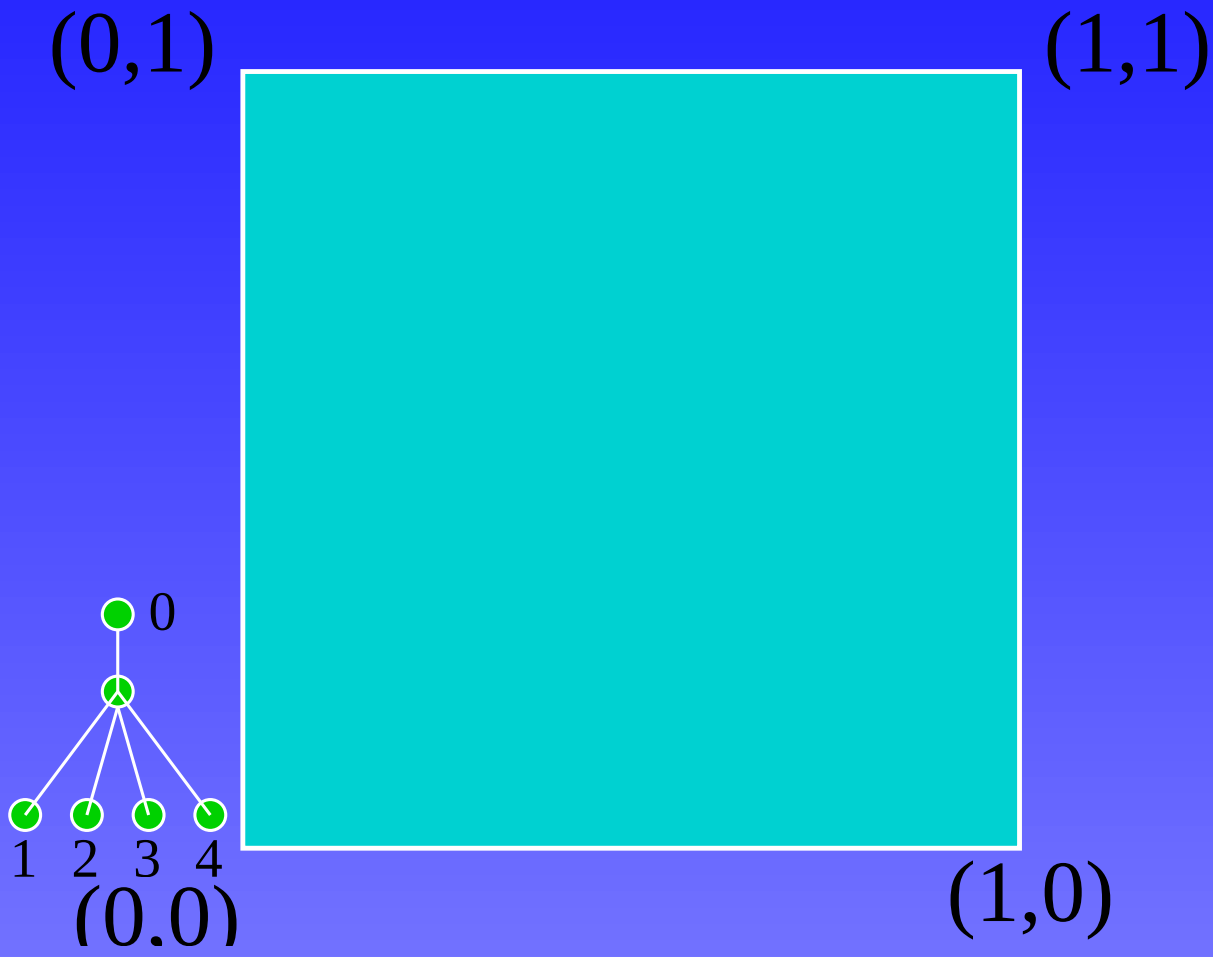


The geometric construction is joint work with Louis Billera and Karen Vogtmann [0]

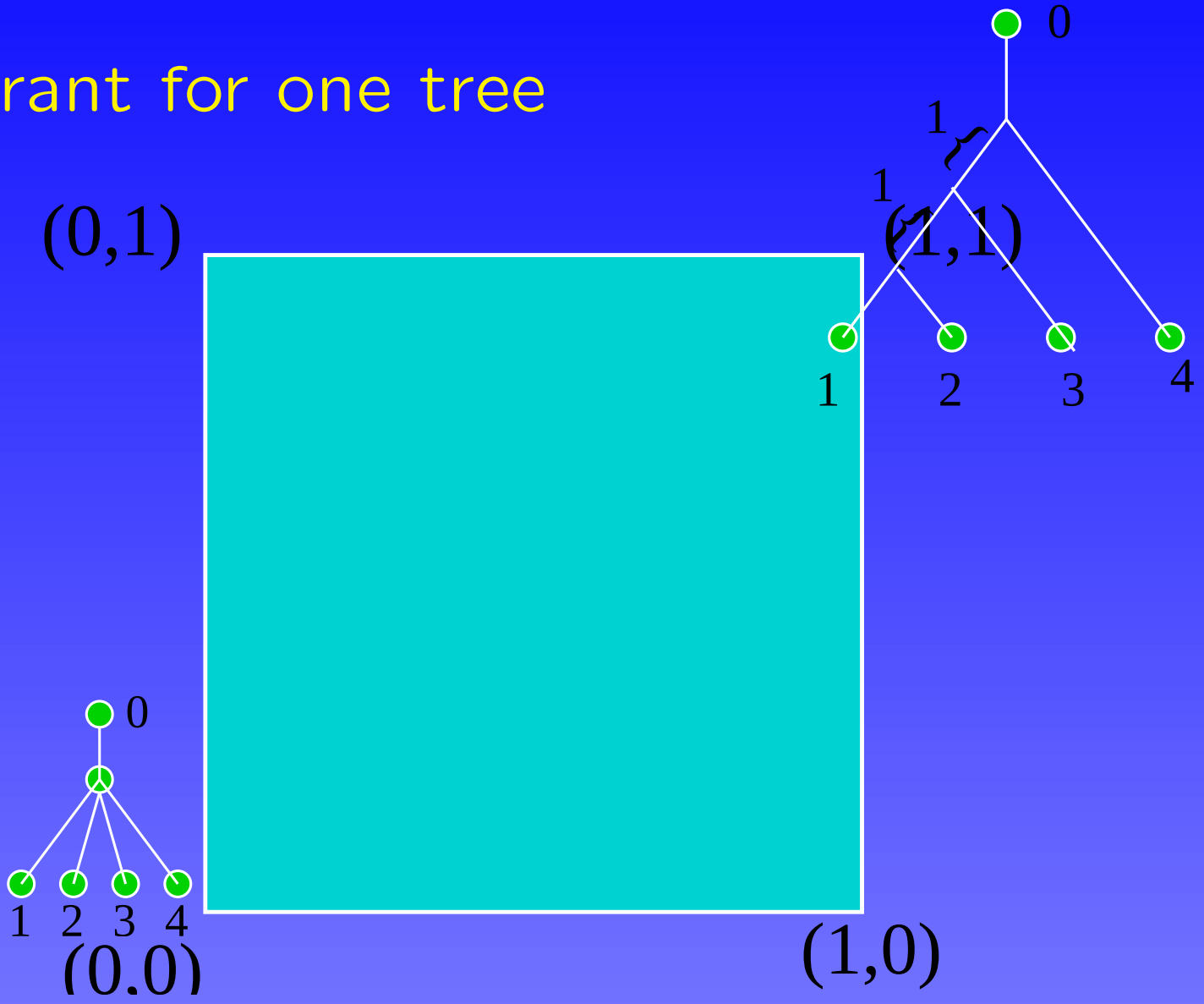
The quadrant for one tree



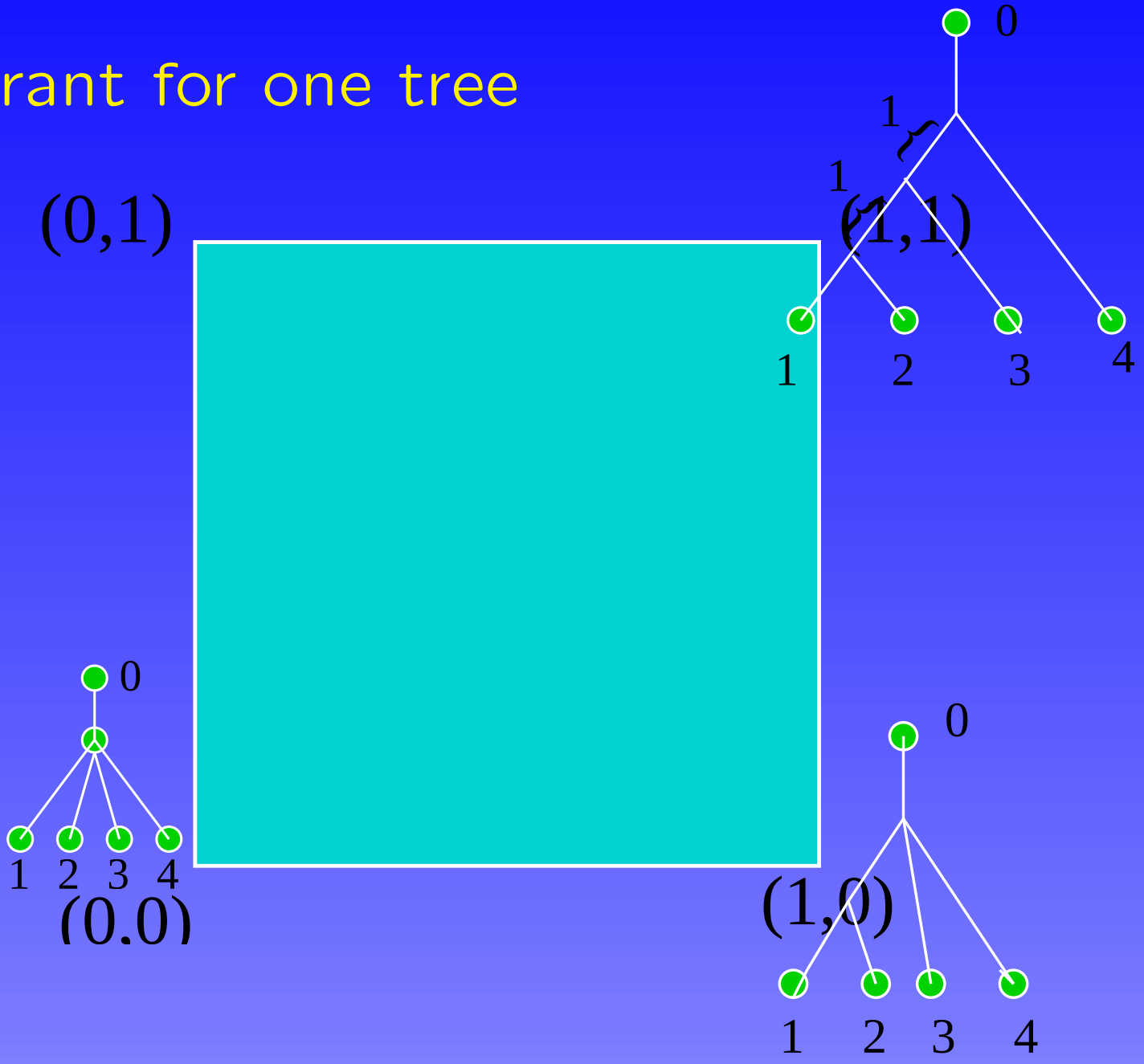
The quadrant for one tree



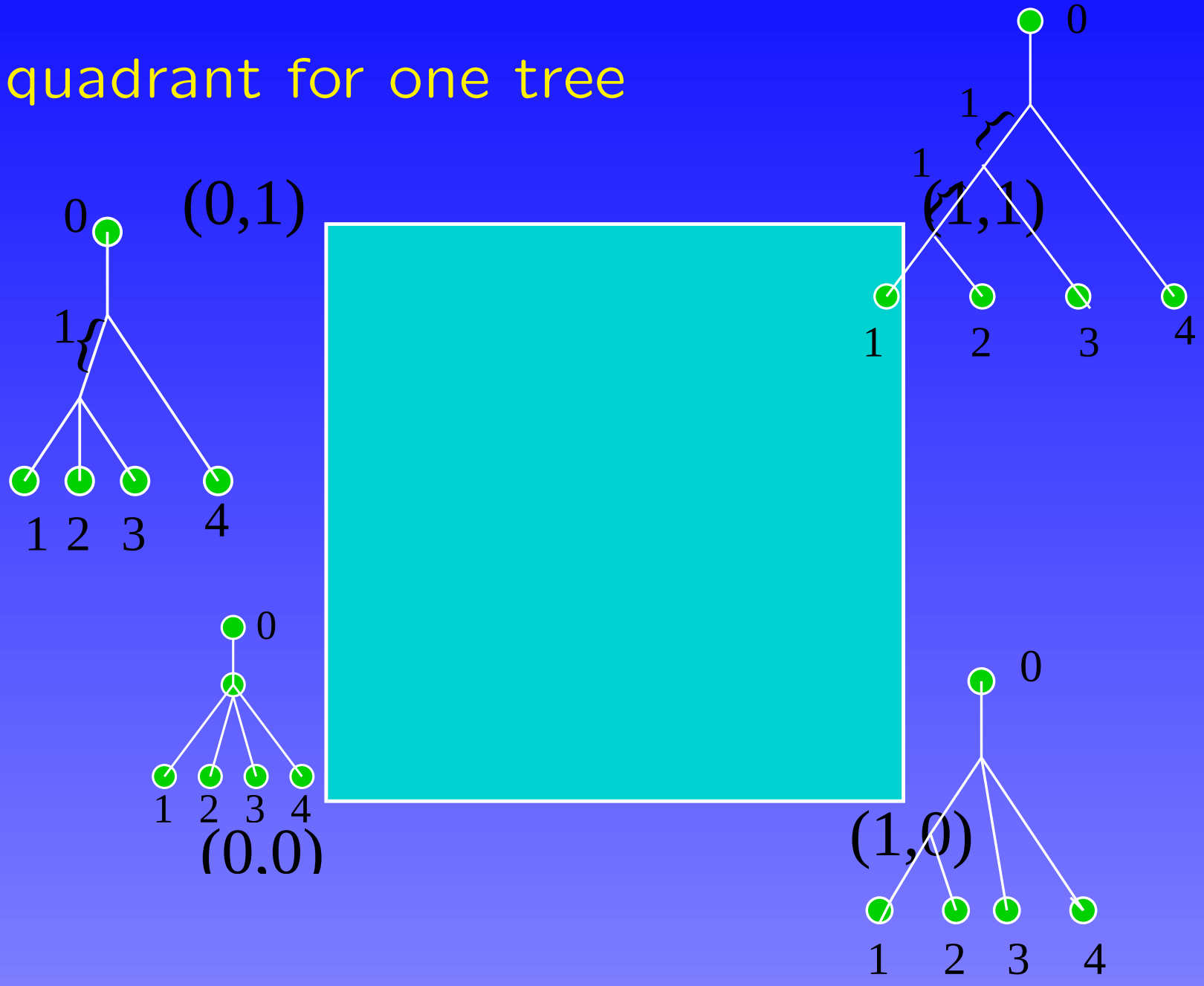
The quadrant for one tree



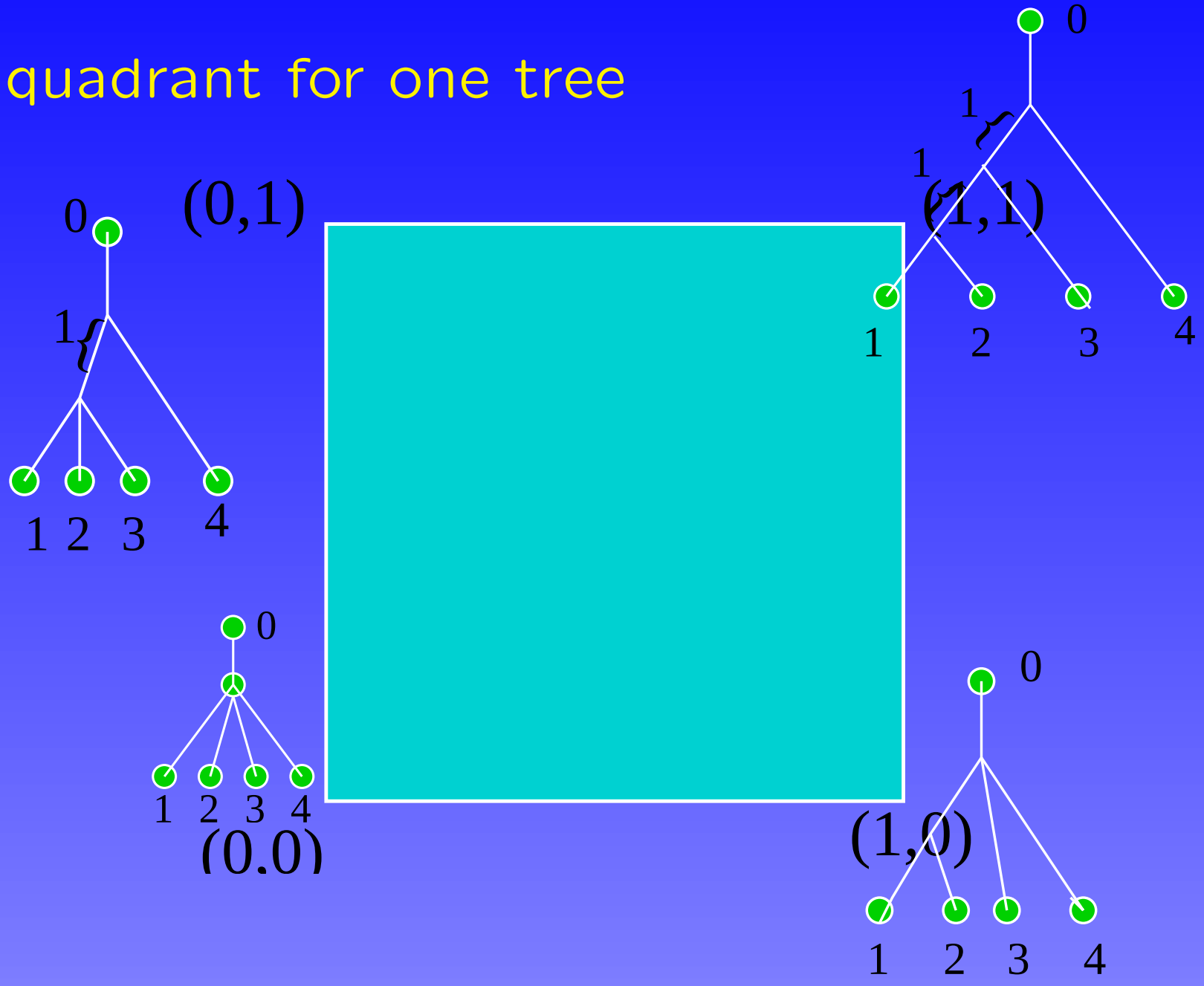
The quadrant for one tree



The quadrant for one tree



The quadrant for one tree



The cube complex

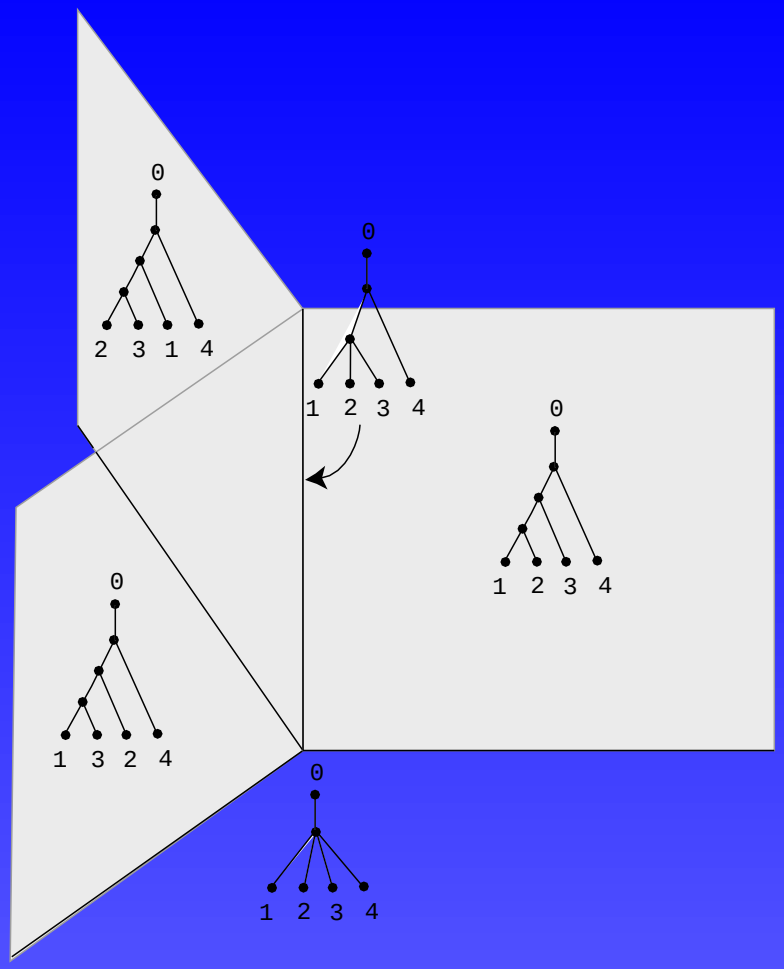
A binary n-tree has the maximal possible number of interior edges $(n-2)$. It determines the largest possible dimensional quadrant which is $n-2$-dimensional. The quadrant corresponding to each tree which is not binary appears as a boundary face of at least three binary trees; in particular the origin of each quadrant corresponds to the (unique) tree with no interior edges.

The cube complex

A binary n-tree has the maximal possible number of interior edges $(n-2)$. It determines the largest possible dimensional quadrant which is $n-2$-dimensional. The quadrant corresponding to each tree which is not binary appears as a boundary face of at least three binary trees; in particular the origin of each quadrant corresponds to the (unique) tree with no interior edges. $\mathcal{T}_n$ is built by taking one $n-2$-dimensional quadrant for each of the $(2n-3)!! = (2n-3)*(2n-5)*\cdots*5*3*1$ possible binary trees, and gluing them together along their common faces.

For $n=3$ there are three binary trees, each with 1 interior edge. Each tree thus determines a 1-dimensional "quadrant," i.e. a ray from the origin. The three rays are identified at their origins. Figure for n=3.

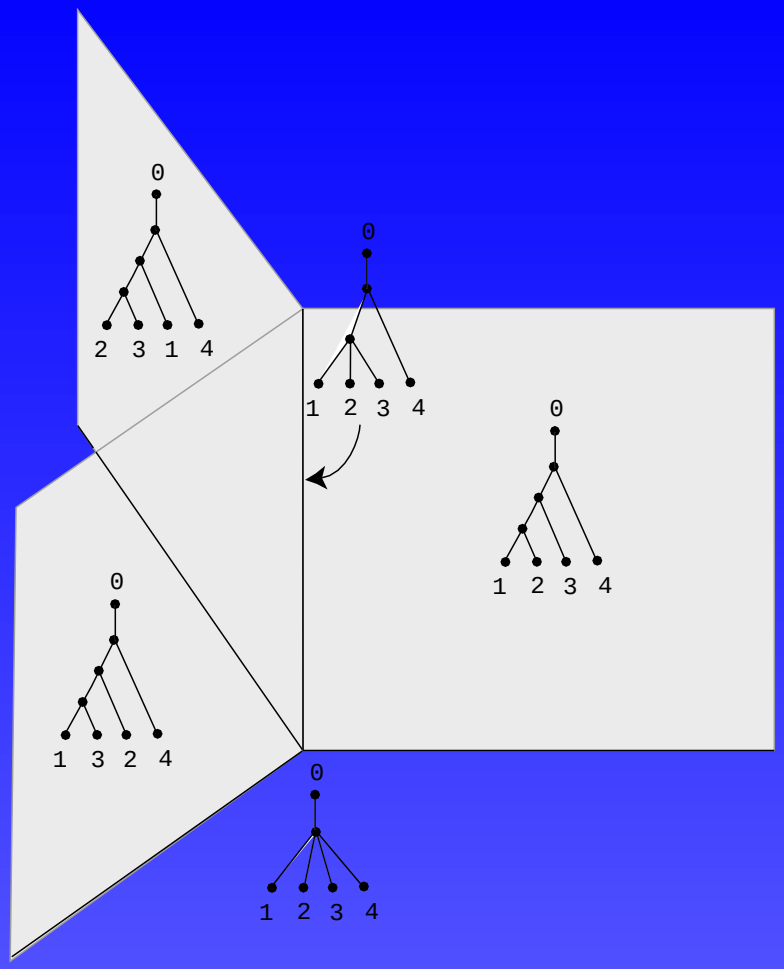
Boundary



Three quadrants sharing a ray for n=4

Note that the bottom boundary rays form a copy of $\mathcal{T}_3$ embedded in $\mathcal{T}_4$.

Boundary

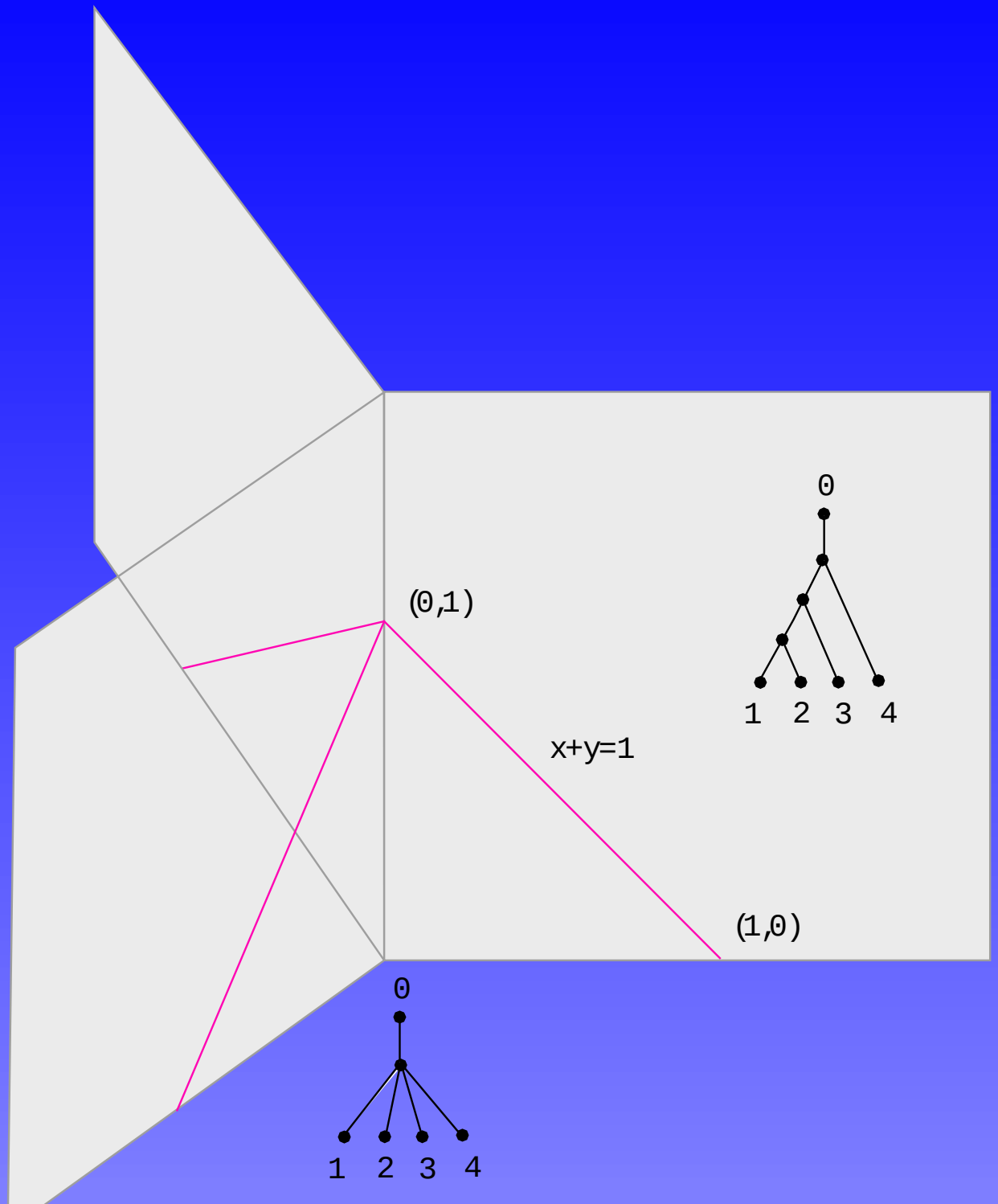


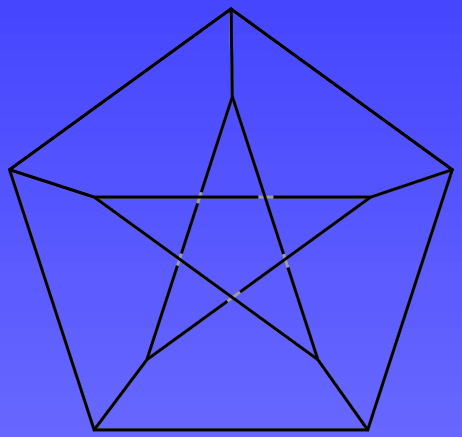
Three quadrants sharing a ray for n=4

Note that the bottom boundary rays form a copy of $\mathcal{T}_3$ embedded in $\mathcal{T}_4$. In general, $\mathcal{T}_n$ contains many embedded copies of $\mathcal{T}_k$ for $k < n$. For example, there is a copy of $\mathcal{T}_k$ for each interior vertex P such that the associated tree T_P has k leaves.

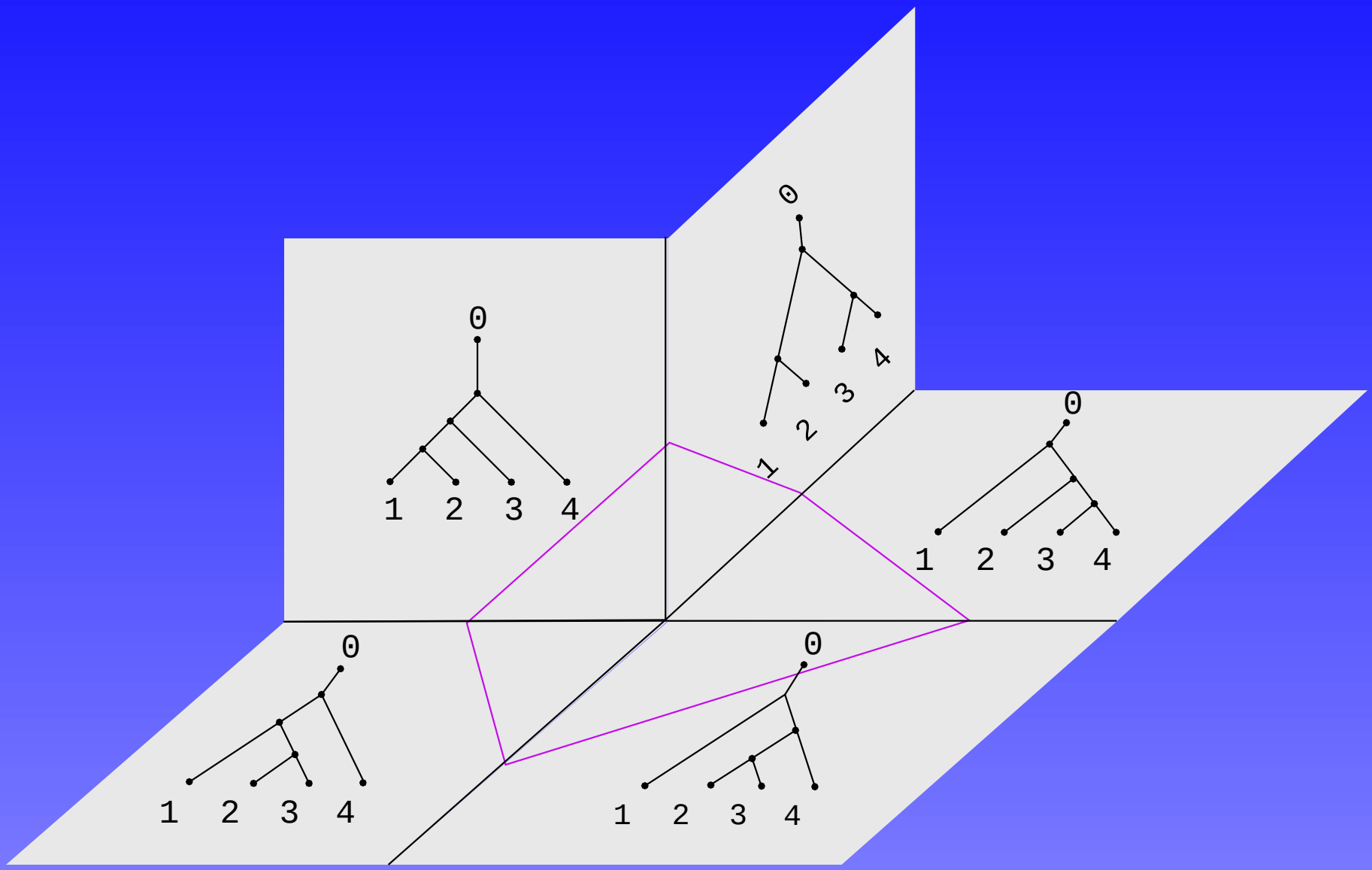
Link to the origin

All 15 quadrants for $n = 4$ share the same origin. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray; this graph is called the *link of the origin*.

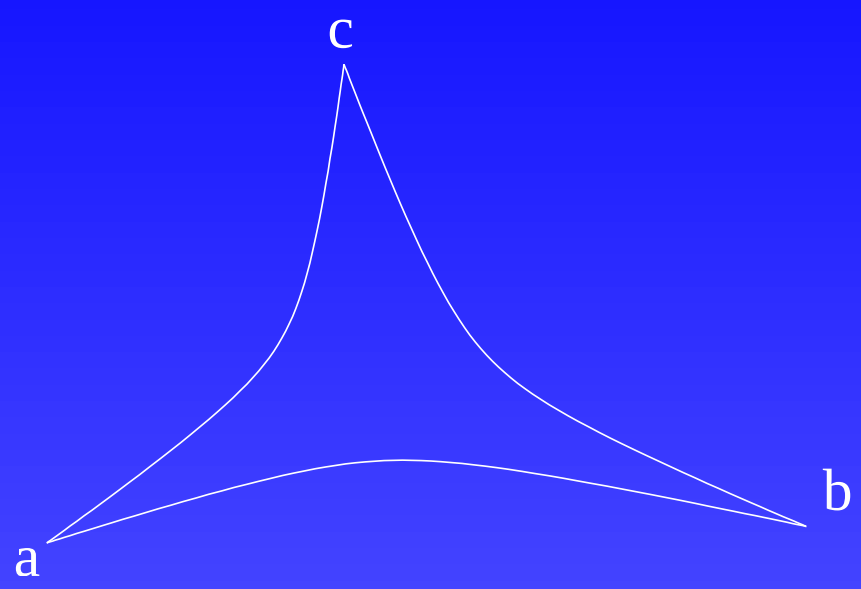

(0,1)
x+y=1
(1,0)
0
1 2 3 4
0
1 2 3 4

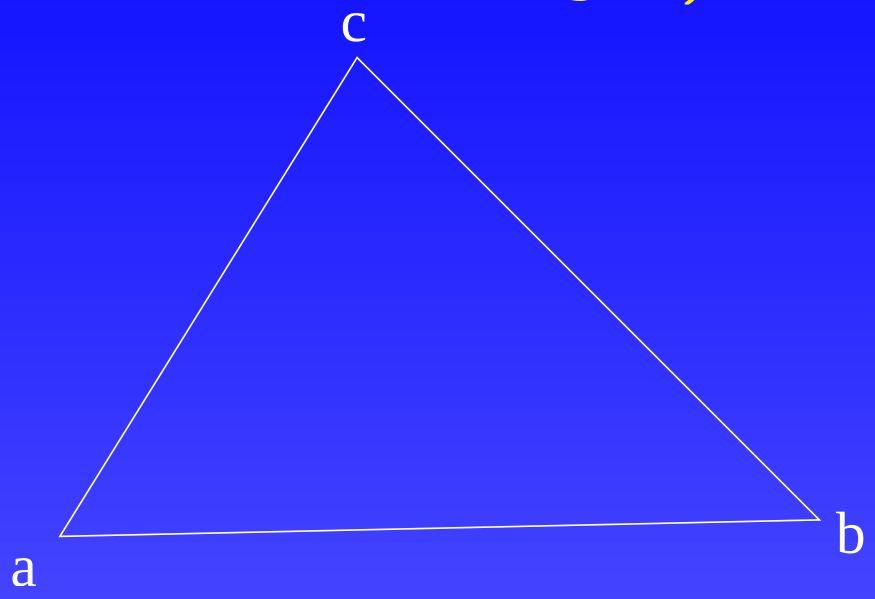


Associahedron



CAT(0) space, (see Bridson and Haefliger)





given any three points a, b and c in X, with distances $d_1 = d(b, c), d_2 = d(a, c)$ and $d_3 = d(a, b)$, form a "comparison triangle" in the Euclidean plane with vertices a', b' and c' with side lengths $d_1 = d(b', c'), d_2 = d(a', c')$ and $d_3 = d(a', b')$. If x is a point on the geodesic from a to b, at distance d from a, find the corresponding point x' on the straight line from a' to b' at distance d from a'. Then $d(x', c') \leq d(x, c)$.

For further reference see Gromov [0].

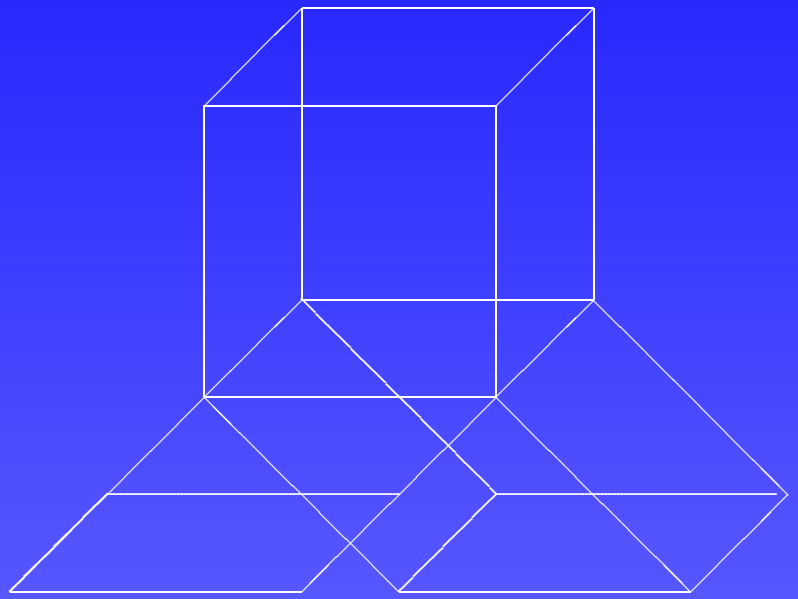
Consequences

- We can define centroids. For two points, take their geodesic and then take the middle of it. Theorem
- We can make convex hulls. $\longrightarrow$ Confidence regions.
- We can use Mallow's model.
- We know the number of neighbors of each tree.

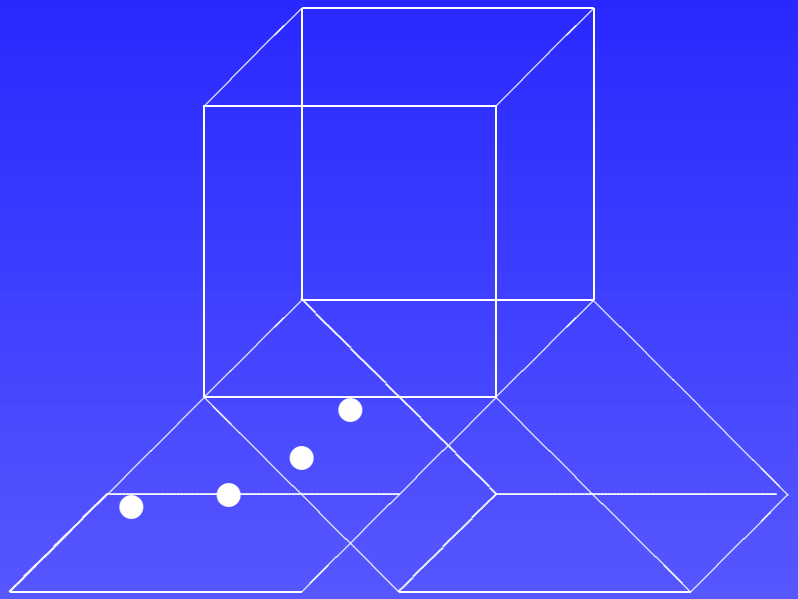
Consequences

- We can define centroids. For two points, take their geodesic and then take the middle of it. Theorem
- We can make convex hulls. $\longrightarrow$ Confidence regions.
- We can use Mallow's model.
- We know the number of neighbors of each tree. $(2(n-2))$

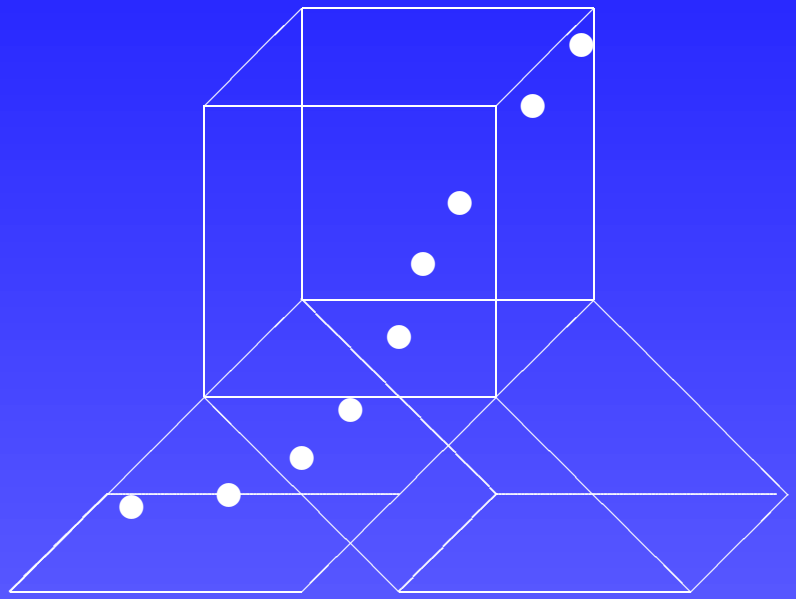
Tree trajectories



Tree trajectories



Tree trajectories



How can abstract mathematics help?

- Decompositions that can be generalisable.
- Geometric Picture of Tree Space
  - A space for comparisons.
  - Ways of *projecting*.
  - *Follow* trees as they change, (paths of trees)
  - Centroids of trees
  - Neighborhoods (convex hulls of trees)....
  - Averages of trees
- Justification of commonsense, ground for generalizations.

Theorem In any CAT(0) space X,

1. Centroids exist for any finite set $Y \subset X$.

2. The centroid function is convex.

Proof: The proof is by induction on $n = |Y|$. The case $n = 2$ is in Bridson and Haefliger, 1999.

Suppose $n \geq 3$ and we have a convex centroid function $c(W)$ for $|W| = n-1$. Let $Y = \{y_1, \ldots, y_n\}$ and $Y_i = Y \setminus \{y_i\}$. Suppose Y has diameter M. Then by convexity for $(n-1)$-sets, $d(c(Y_i), c(Y_j)) \leq \frac{1}{n-1}\, d(y_i, y_j) \leq \frac{1}{n-1} M$, and so diam $c^1(Y) \leq \frac{1}{n-1} M$. Thus the diameter of $c^k(Y)$ is bounded above by $\left(\frac{1}{n-1}\right)^k M$ and so goes to zero.

To show convergence, let M_k denote the diameter of $c^k(Y)$, and consider the sequence $z_k \in c^k(Y)$, where $z_0 = y_1$, $z_1 = c(Y_1)$, $z_2 =$

$c(\{c(Y_i) : i \neq 1\})$, etc. It follows by convexity for $(n-1)$-sets that $d(z_k, z_{k+1}) \leq \frac{1}{n-1} M_k$. Thus for $l \geq k$,

$$d(z_k, z_l) \leq M_k + \frac{1}{n-1}\, M_k + \left(\frac{1}{n-1}\right)^2 M_k + \cdots \;=\; \frac{n-1}{n-2}\, M_k,$$

showing that $\{z_k\}$ is a Cauchy sequence. Thus, centroids exist for n-sets.

To show convexity of $c(Y)$, $|Y| = n$, suppose $Y = \{y_1, \ldots, y_n\}$ and $Y' = \{y'_1, \ldots, y'_n\}$. If $Y_i = Y \setminus \{y_i\}$ and $Y'_i = Y' \setminus \{y'_i\}$, then by convexity for $(n-1)$-sets,

$$d(c(Y_i), c(Y'_i)) \leq \frac{1}{n-1} \sum_{j \neq i} d(y_j, y'_j) \tag{1}$$

for each i. Let $\delta_i = d(y_i, y'_i)$ and $d_0 = (\delta_1, \ldots, \delta_n)$. Then if d_k is the corresponding vector of distances between elements of $c^k(Y)$ and $c^k(Y')$, it follows from (1) that $d_k \leq B_n^k d_0$, where $B_n = \frac{1}{n-1}(J_n - I_n)$, J_n is the $n \times n$ matrix of $1's$ and I_n is the $n \times n$ identity matrix. Since $B_n^k \to \frac{1}{n} J$ as $k \to \infty$, it follows that $d(c(Y), c(Y')) \leq \frac{1}{n} \sum d(y_i, y'_i)$ as desired. back

Y. Amit and D. Geman, *Quantization and recognition with randomized trees*, Neural Computation, 9 (1997), pp. 1545–1588.

L. Billera, S. Holmes, and K. Vogtmann, *Geometry of tree space*, to appear, Statistics, Stanford, 1999.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth, 1984.

M. R. Bridson, *Geodesics and curvature in metric simplicial complexes*, in Group theory from a geometrical viewpoint (Trieste, 1990), World Sci. Publishing, River Edge, NJ, 1991, pp. 373–463.

M. Charleston, *Landscape of trees*, `http://taxonomy.zoology.gla.ac.uk/~mac/landsca/trees.html.`, ?? (1996), pp. –.

B. Charnomordic and S. Holmes, `Dnaview`, *an interactive viewer for alignment and tree building*, unpublished software, (1997).

P. Diaconis, *Group Representations in Probability and Statistics*, Institute of Mathematical Statistics, 1988.

—, *A generalization of spectral analysis with application to ranked data*, The Annals of Statistics, 17 (1989), pp. 949–979.

P. Diaconis and S. P. Holmes, *Gray codes and randomization procedures*, Statistics and Computing, (1994), pp. 287–302.

P. W. Diaconis and S. P. Holmes, *Matchings and phylogenetic trees*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 14600–14602 (electronic).

—, *Matchings and phylogenetic trees*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 14600–14602 (electronic).

B. Efron, E. Halloran, and S. P. Holmes, *Bootstrap confidence levels for phylogenetic trees*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 13429–34.

B. Efron and R. Tibshirani, *The problem of regions*, Ann. Statist., 26 (1998), pp. 1687–1718.

M. Eisen, P. Spellman, P. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns.*, Proc Natl Acad Sci USA, (1998).

J. Friedman, *Greedy function approximation: A gradient boosting machine*, tech. rep., Stanford Statistics Dept., 1999. http://www-stat.stanfor

M. Gromov, *Hyperbolic groups*, in Essays in group theory, Springer, New York, 1987, pp. 75–263.

K. Hayasaka, T. Gojobori, and S. Horai, *Molecular phylogeny and evolution of primate mitochondrial dna.*, Mol. Biol. Evol., 5/6 (1988), pp. 626–644.

S. Holmes, *Phylogenies: An overview*, in Statistics and Genetics, E. Halloran and S. Geisser, eds., no. 81 in IMA, Springer Verlag, NY, 1999.

S. Holmes and P. Diaconis, *Computing with Trees*, Interface Foundation of North America, 1999.

S. Li, D. K. Pearl, and H. Doss, *Phylogenetic tree construction using mcmc*, Journ. American Statistical Association (to appear)., (2000). Ohio Statistics Dept., ().

B. MacFadden and R. Hulbert Jr, *Explosive speciation at the base of the adaptive radiation of miocene grazing horses*, Nature, 336 (1988), pp. 466–468.

B. Mau, M. A. Newton, and B. Larget, *Bayesian phylogenetic inference via markov chain monte carlo methods*, Biometrics, (1999).

E. Schröder, Zeit. für. Math. Phys., 15 (1870), pp. 361–376.

W. Shannon and D. Banks, *Combining classification trees using maximum likelihood estimation*, Statistics In Medicine, 18 (1999), pp. 727–740.

G. L. Thompson, *Generalized permutation polytopes and exploratory graphical methods for ranked data*, The Annals of Statistics, 21 (1993), pp. 1401–1430.

G. M. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995.