Confidence Statements for Phylogenetic Trees

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

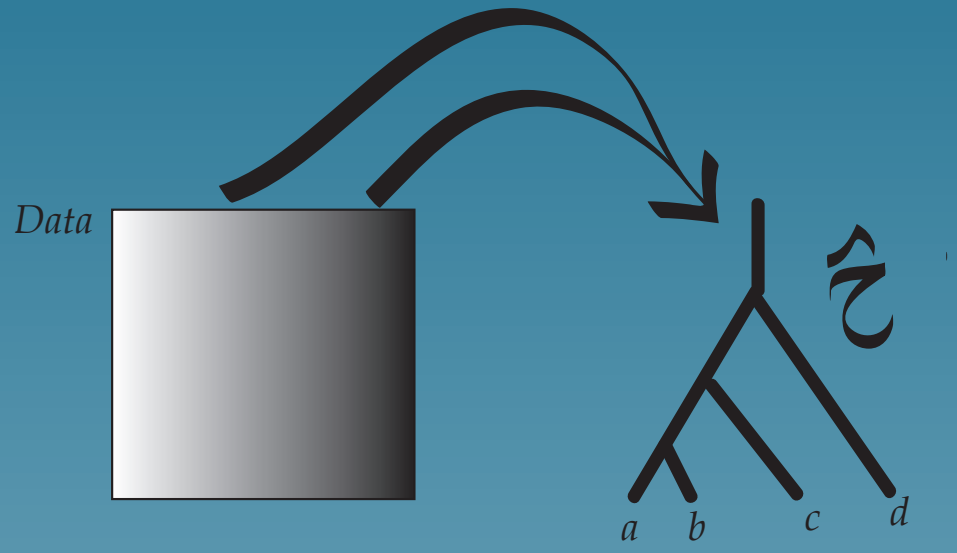
Funded in part by a grant from NSF-DMS
Collaborators: Karen Vogtmann, Persi Diaconis and Lou Billera, Aaron Staple, Henry Towsner.

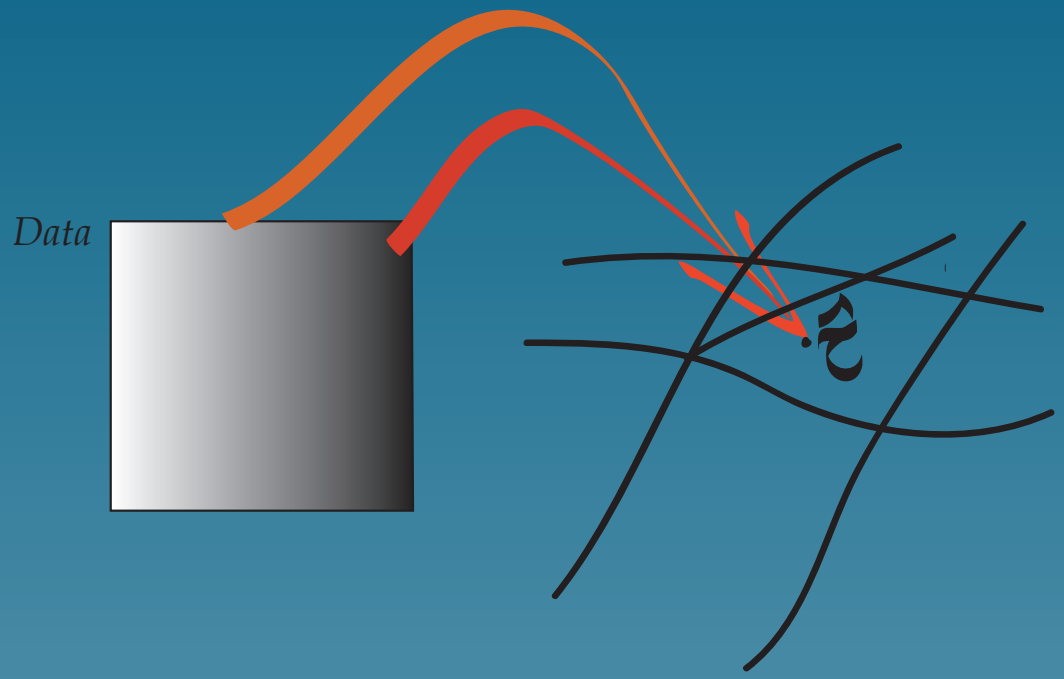
Outline

- The statistical paradigm.
  - ⋆ Estimates and confidence.
  - ⋆ Statistical approaches to variability.
  - ⋆ Robustness
  - ⋆ Inspirations from ranked data.
- Building a treespace with a natural distance.
- The Bootstrap.
- Special cases: multivariate statistics and Treespace

Estimation in Treespace

Estimate $\hat{\tau}$ computed from the data:




Data

Data can be:

- Binary

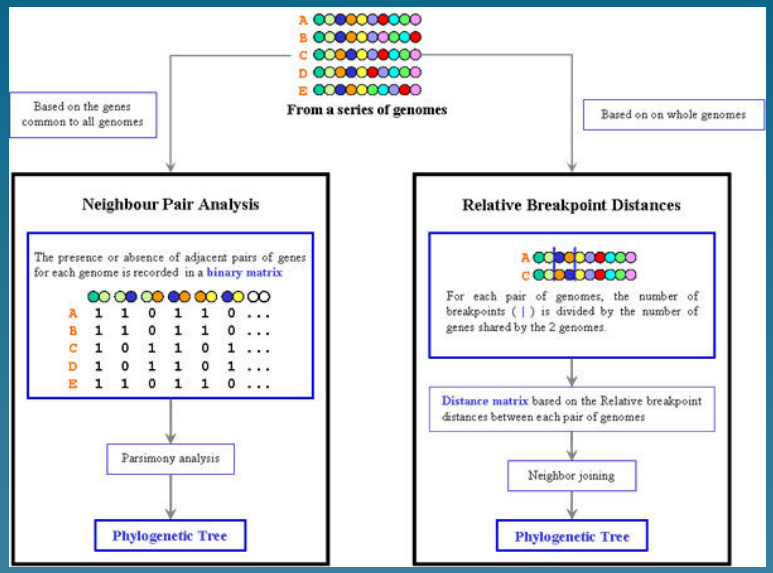
```
Lemur_cat   00000000000001010100000
Tarsius_s   10000010000000010000000
Saimiri_s   10000010000001010000000
Macaca_sy   00000000000000010000000
Macaca_fa   10000010000000010000000
```

- Aligned

DNA Data for 12 species of primates Mitochondria, 898 characters on 12 species,(Hayasaka, K., T. Gojobori, and S. Horai. 1988).

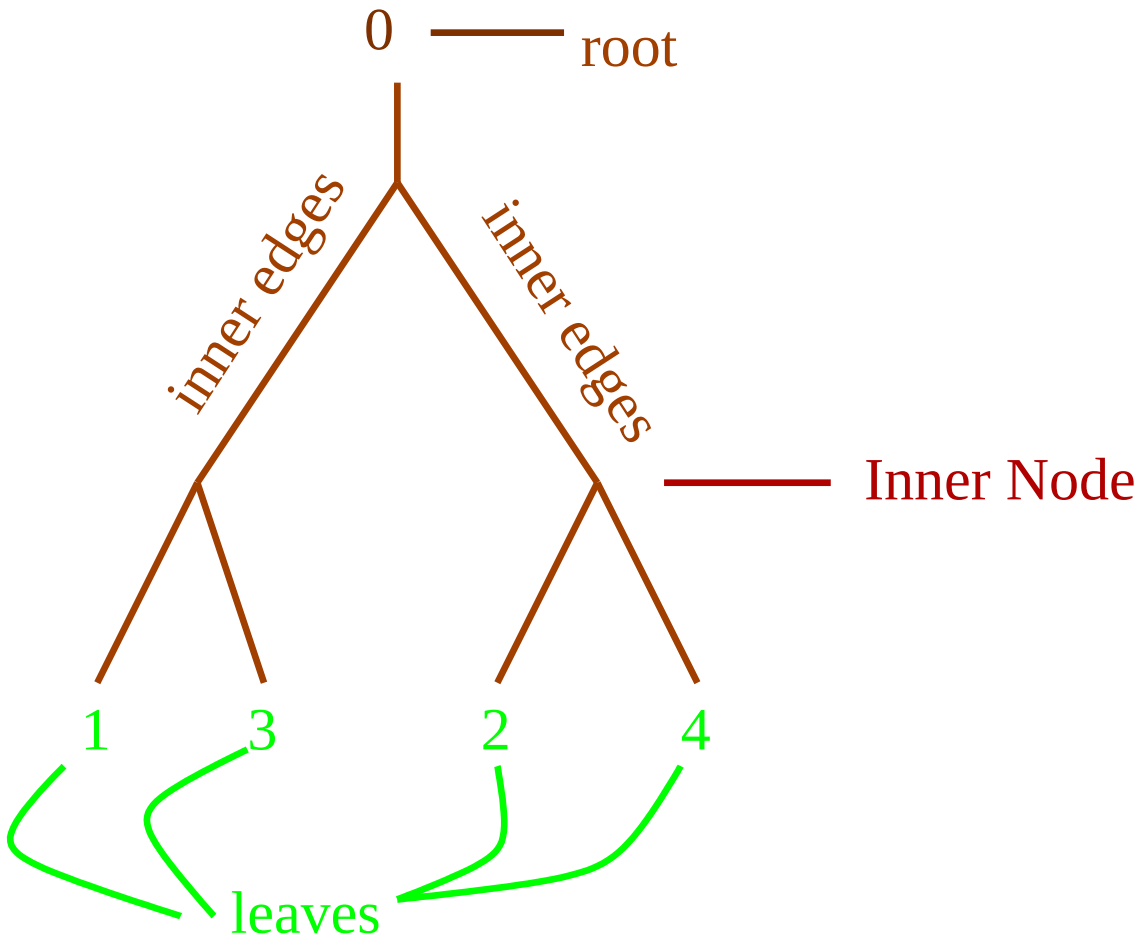
```
  12  60
Lemur_cat AAGCTTCATA GGAGCAACCA TTCTAATAAT CGCACATGGC CTTACATCAT CCATATTATT
Tarsius_s AAGTTTCATT GGAGCCACCA CTCTTATAAT TGCCCATGGC CTCACCTCCT CCCTATTATT
Saimiri_s AAGCTTCACC GGCGCAATGA TCCTAATAAT CGCTCACGGG TTTACTTCGT CTATGCTATT
Macaca_sy AAGCTTCTCC GGTGCAACTA TCCTTATAGT TGCCCATGGA CTCACCTCTT CCATATACTT
Macaca_fa AAGCTTCTCC GGCGCAACCA CCCTTATAAT CGCCCACGGG CTCACCTCTT CCATGTATTT
Macaca_mu AAGCTTTTCT GGCGCAACCA TCCTCATGAT TGCTCACGGA CTCACCTCTT CCATATATTT
```

- Gene order



These data sets usually come with their own metrics.

The parameter is a semi-labeled binary Tree



Statistical Paradigms

Classical Frequentist

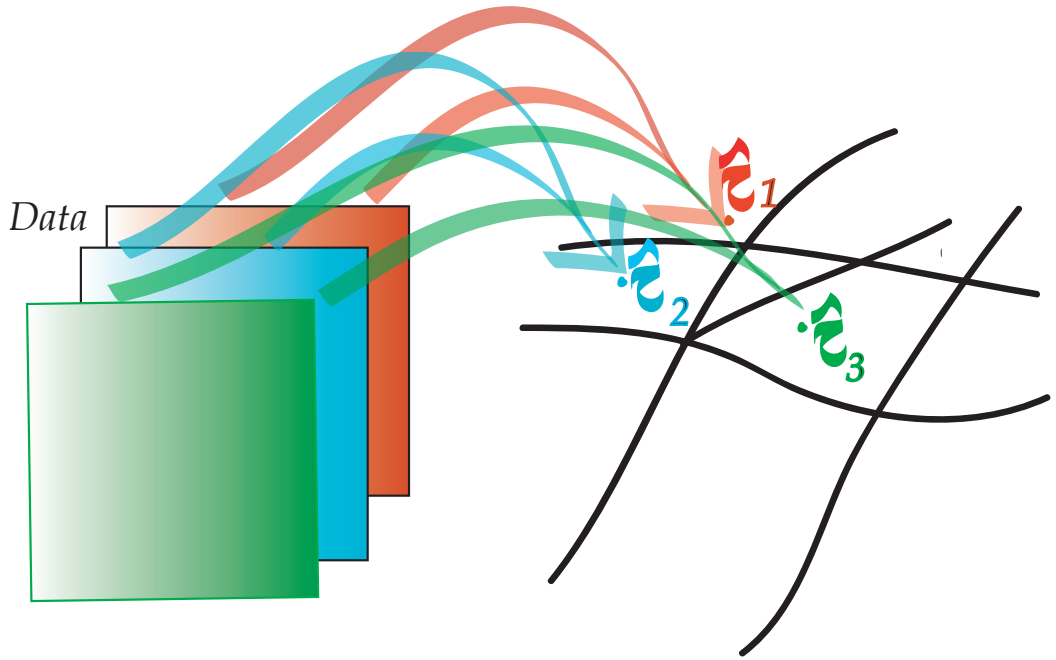
- estimate the parameter,
  (either in a parametric (ML) way,
  semiparametric (Distance based methods),
  or nonparametric way (Parsimony))
- find the sampling Distribution of the estimator.

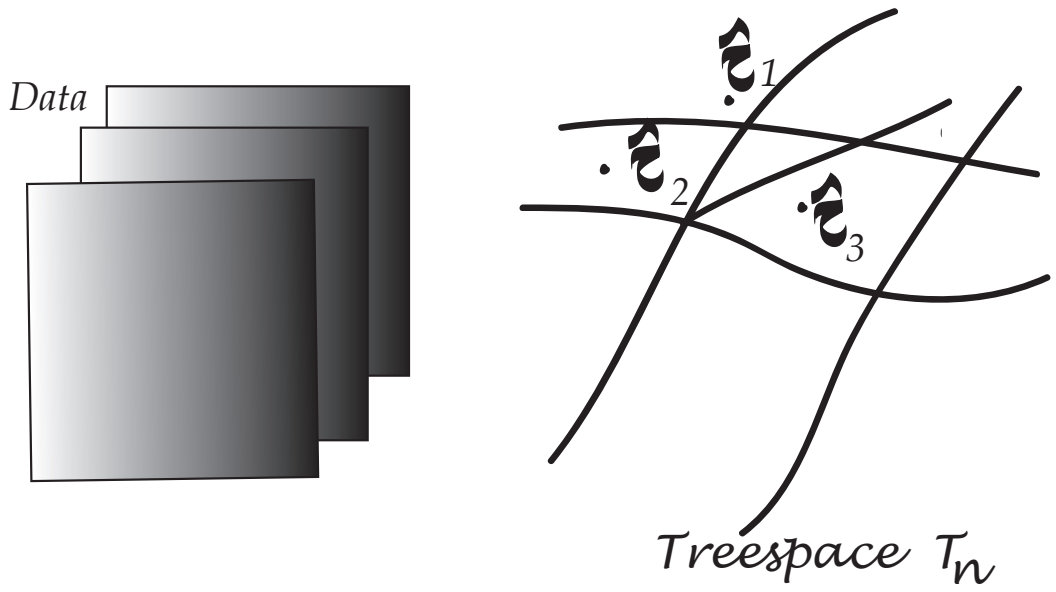
Bayesian

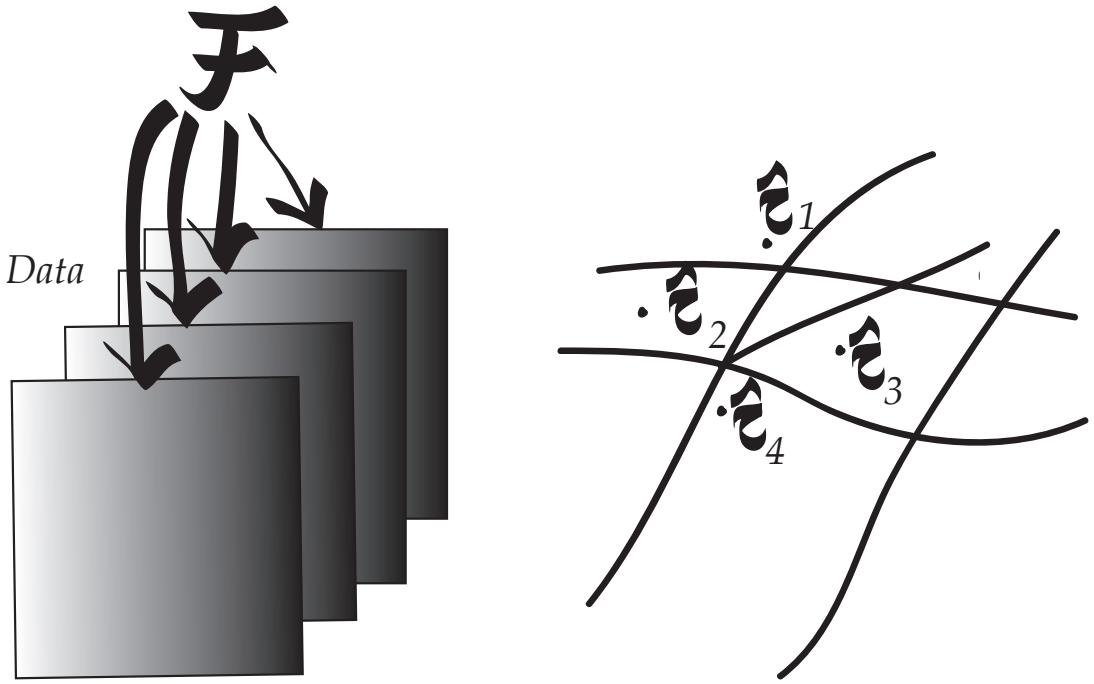
- Specify a Prior Distribution
- Update the prior using the Data
- Compute the Posterior Distribution

Difficulties arise as the estimators lie in a non Euclidean space.

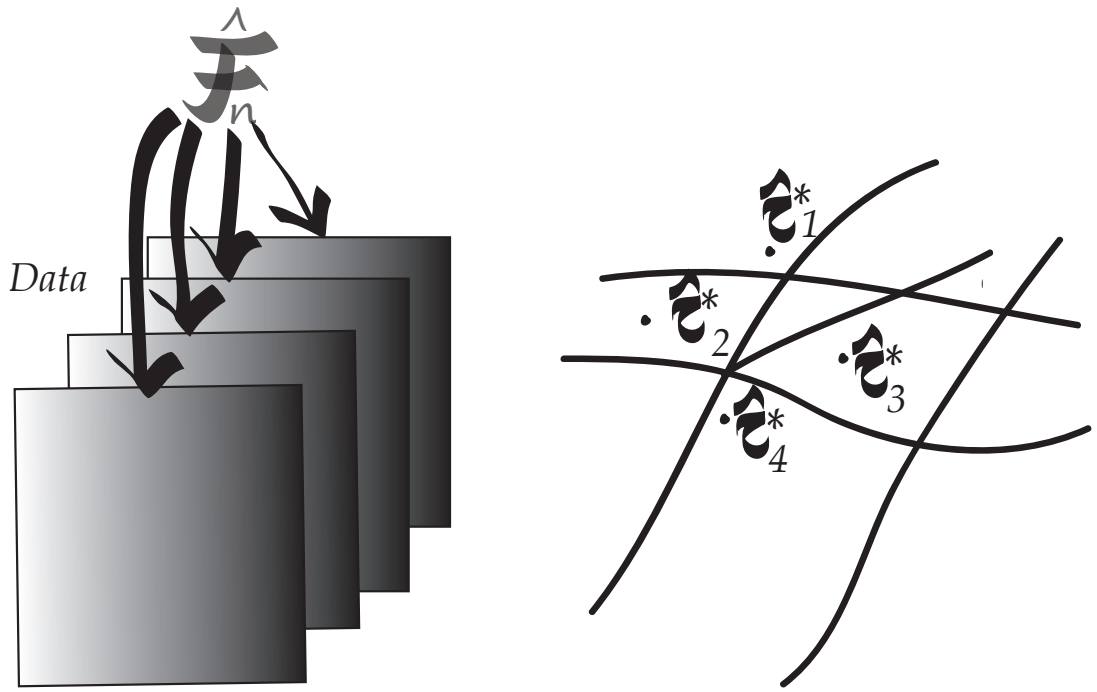
Sampling Distribution for Trees


Data
$\hat{r}_1$
$\hat{r}_2$
$\hat{r}_3$

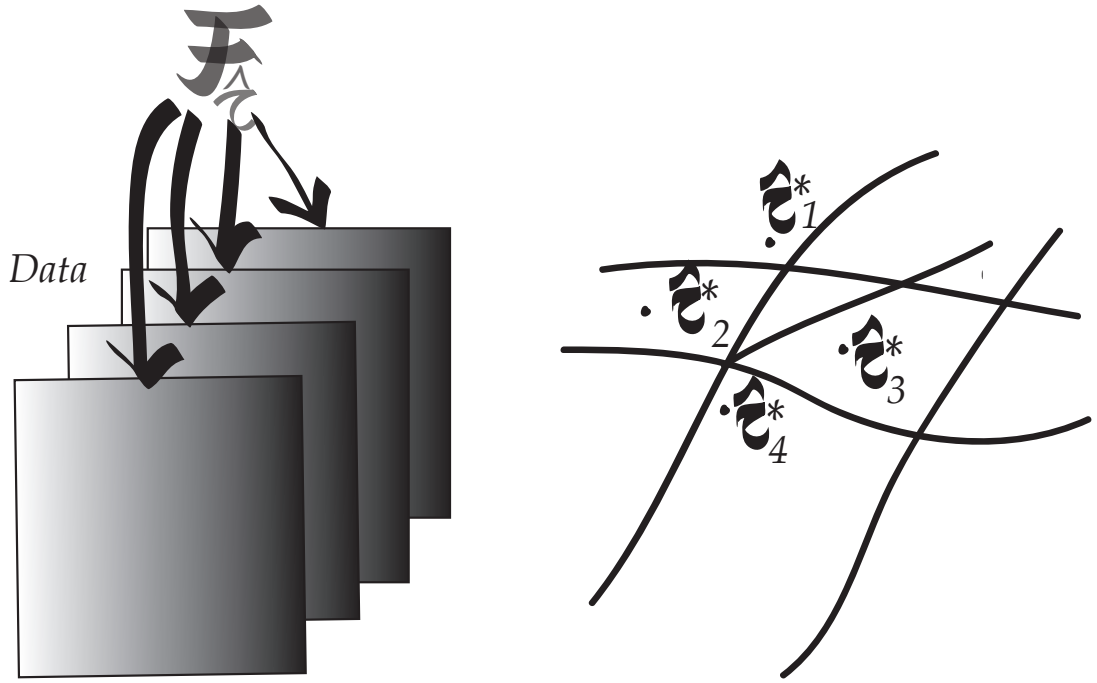

Data
$\hat{\tau}_1$
$\hat{\tau}_2$
$\hat{\tau}_3$
Treespace T_n



True Sampling Distribution



Bootstrap Sampling Distribution
(non parametric)



Bootstrap Sampling Distribution
(parametric)

How do we define distributions on Treespace

- Not the uniform distribution.
- By inspiration from ranked Data: Mallow's model (1957)

$$P(\tau_i) = Ke^{-\lambda d(\tau_i, \tau_0)}$$

  - ⋆ Uses a central tree τ_0.
  - ⋆ Uses a distance d in treespace.
  - ⋆ But very symmetrical, maybe need a mixture model.

- Other distributions (see Aldous, 2001), one might want to include information about the estimation method used as this influences the shape of the tree.

Classical Statistical Summaries

- Expectation (center of the distribution).
  Open question: What distribution is the consensus such a center of?

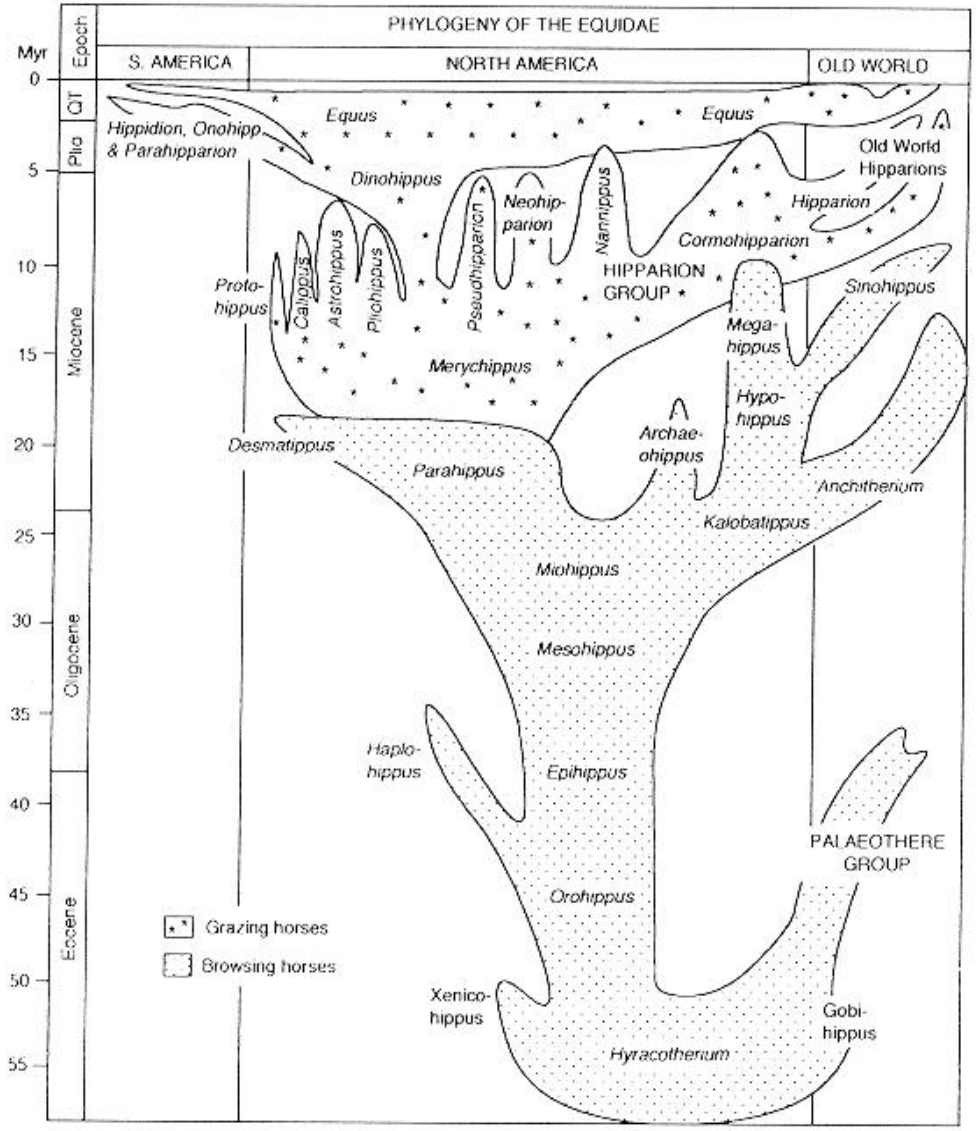
- Median (multivariate median (Tukey, 1972).

- Variance (second moment $E_{P_n} d^2(\hat{\tau}, \tau)$).

- Presence/Absence of a clade.

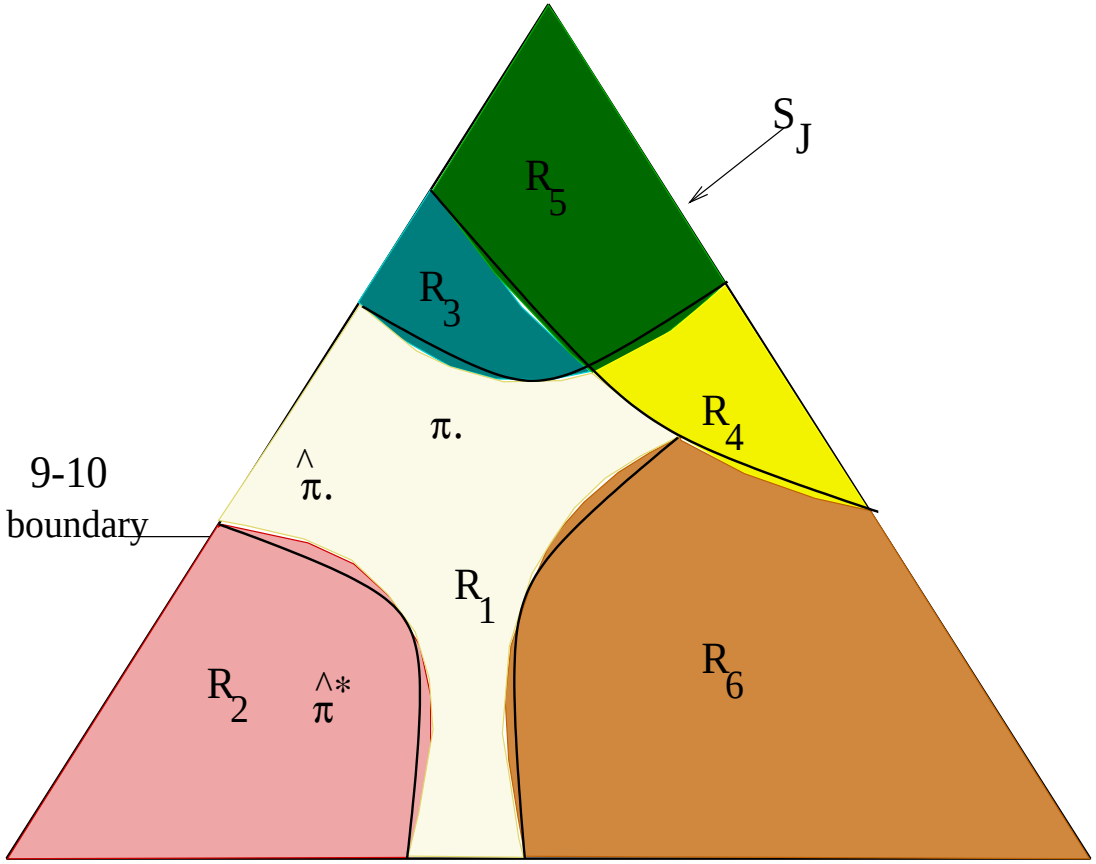
- Summaries of Multivariate Variabilities. (PCA,MDS,..).

Confidence Statements for trees



Confidence Statements in Statistics

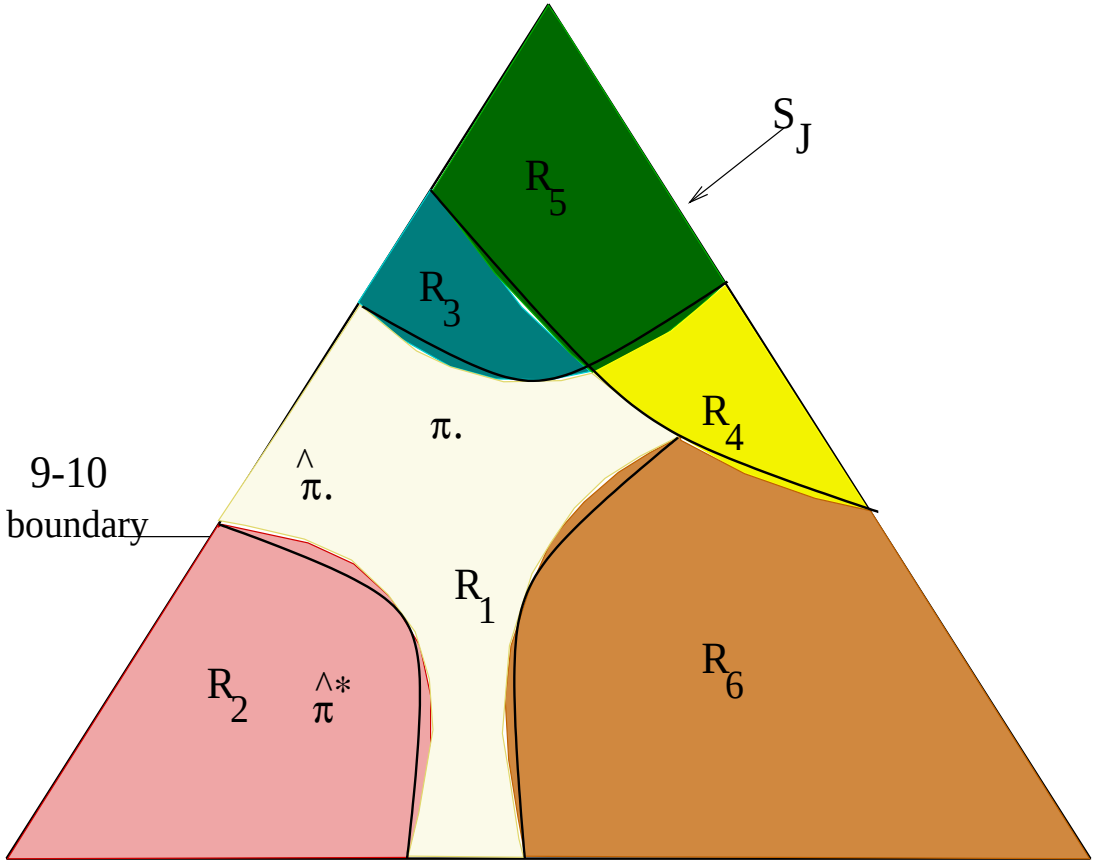
Depend on local and global properties of a neighborhood.



From Efron, Halloran, Holmes, (1996)

Confidence Statements in Statistics

Depend on local and global properties of a neighborhood.



What is the curvature of the boundary?
How many neighbors does a region have?

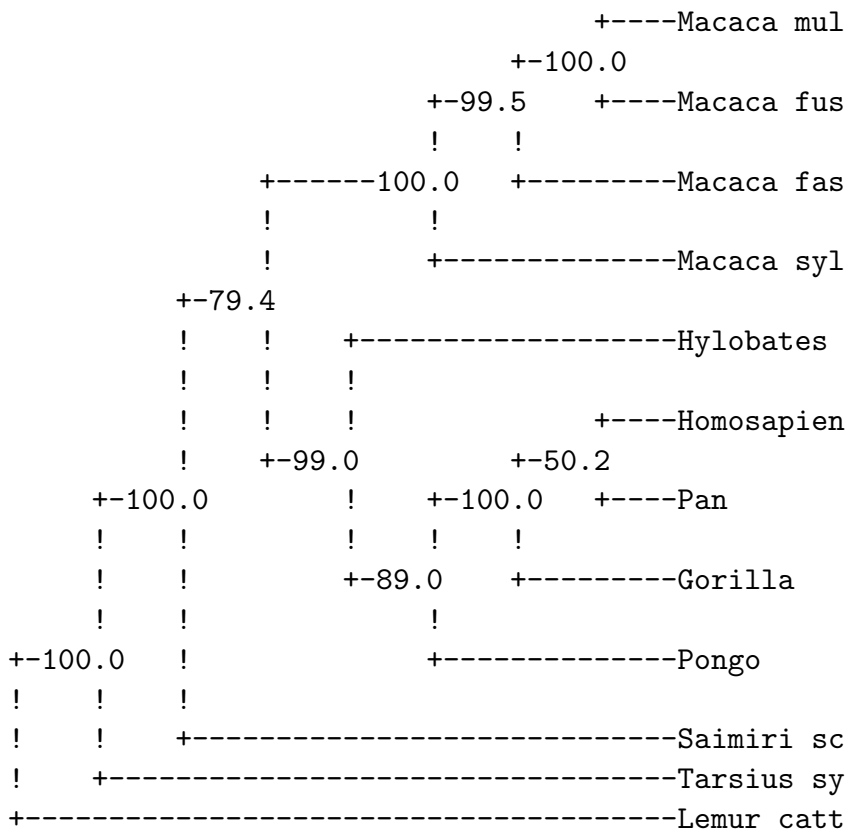
From Efron, Halloran, Holmes, (1996)

Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.

Simple confidence values

- Univariate.
- Multiple Testing.
- Composite Statements.



Do we care about confidence statements for phylogenetic trees?

Cetacees: recognising what is being sold as Whale meat in Japan?

Steve Palumbi, Harvard. Scott Baker, Auckland.

Whale www.DNA.surveillance Earth Trust Press Release

Frequentist Confidence Regions

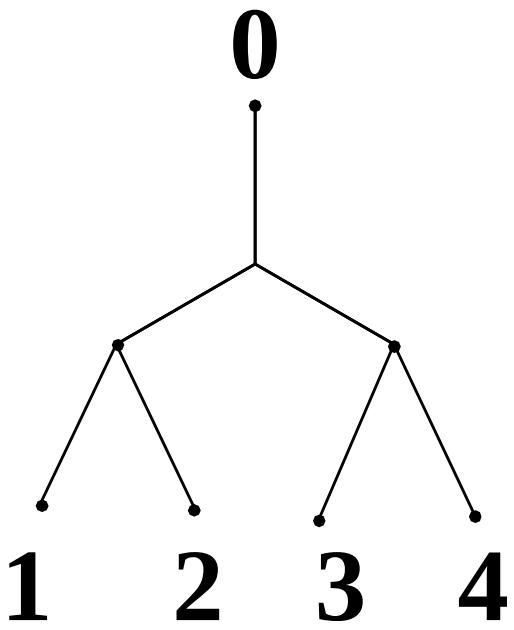
$$P(\tau \in \mathcal{R}_\alpha) = 1 - \alpha$$

We will use the nonparametric approach of Tukey who proposed peeling convex hulls to construct successive 'deeper' confidence regions. But we need a geometrical space to build these regions in.

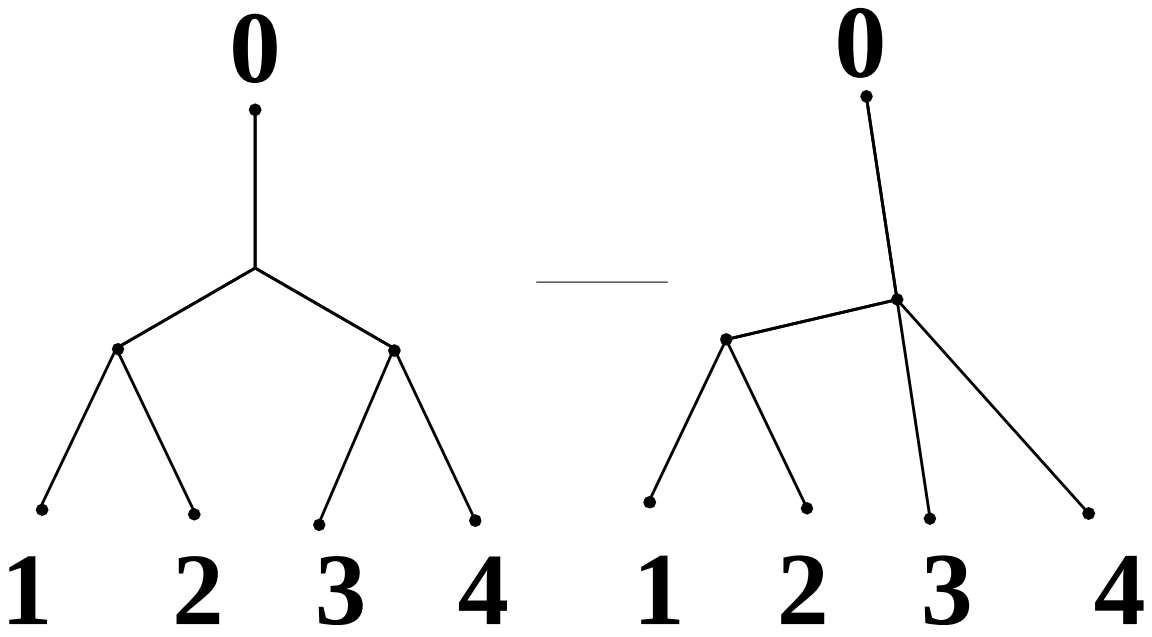
Aims

- Fill Tree Space and make meaningful boundaries.
- Define distances between trees.
- Define neighborhoods, meaningful measures.
- Principal directions of variations in tree space, summarizing : structure + noise.
- Confidence statements, convex hulls.

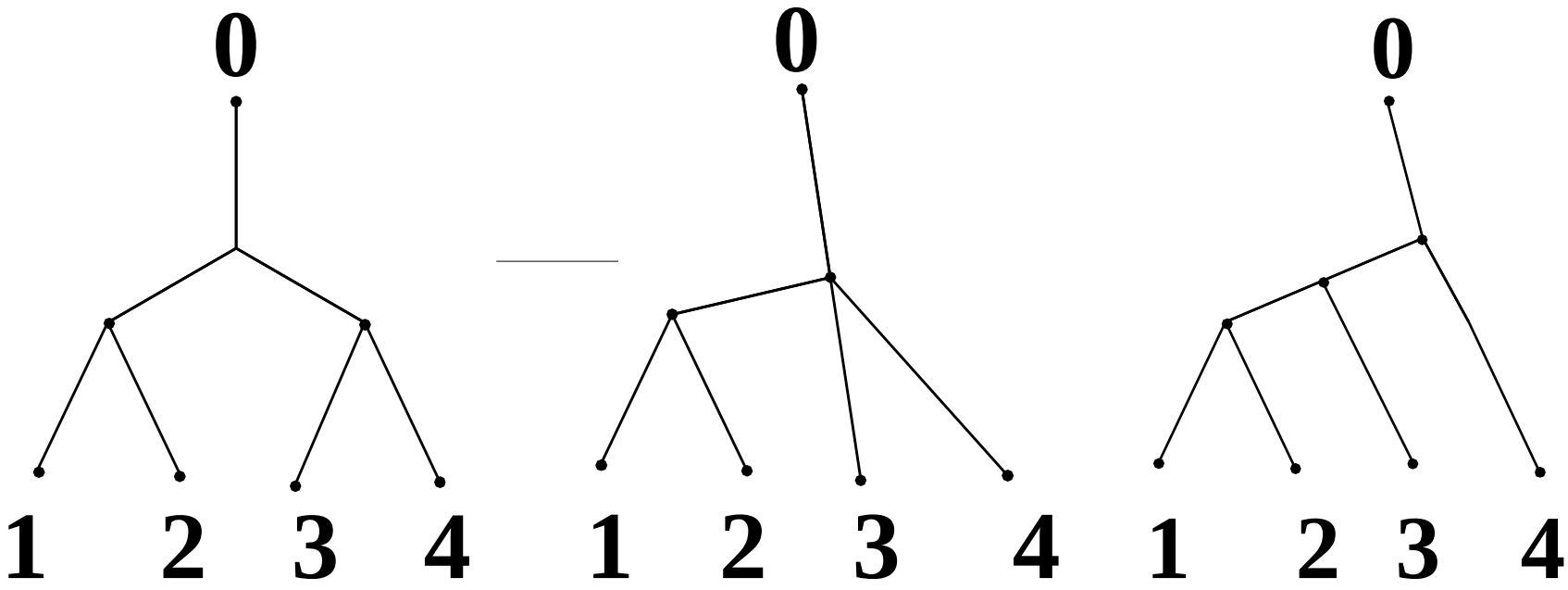
Rotation Moves



Rotation Moves

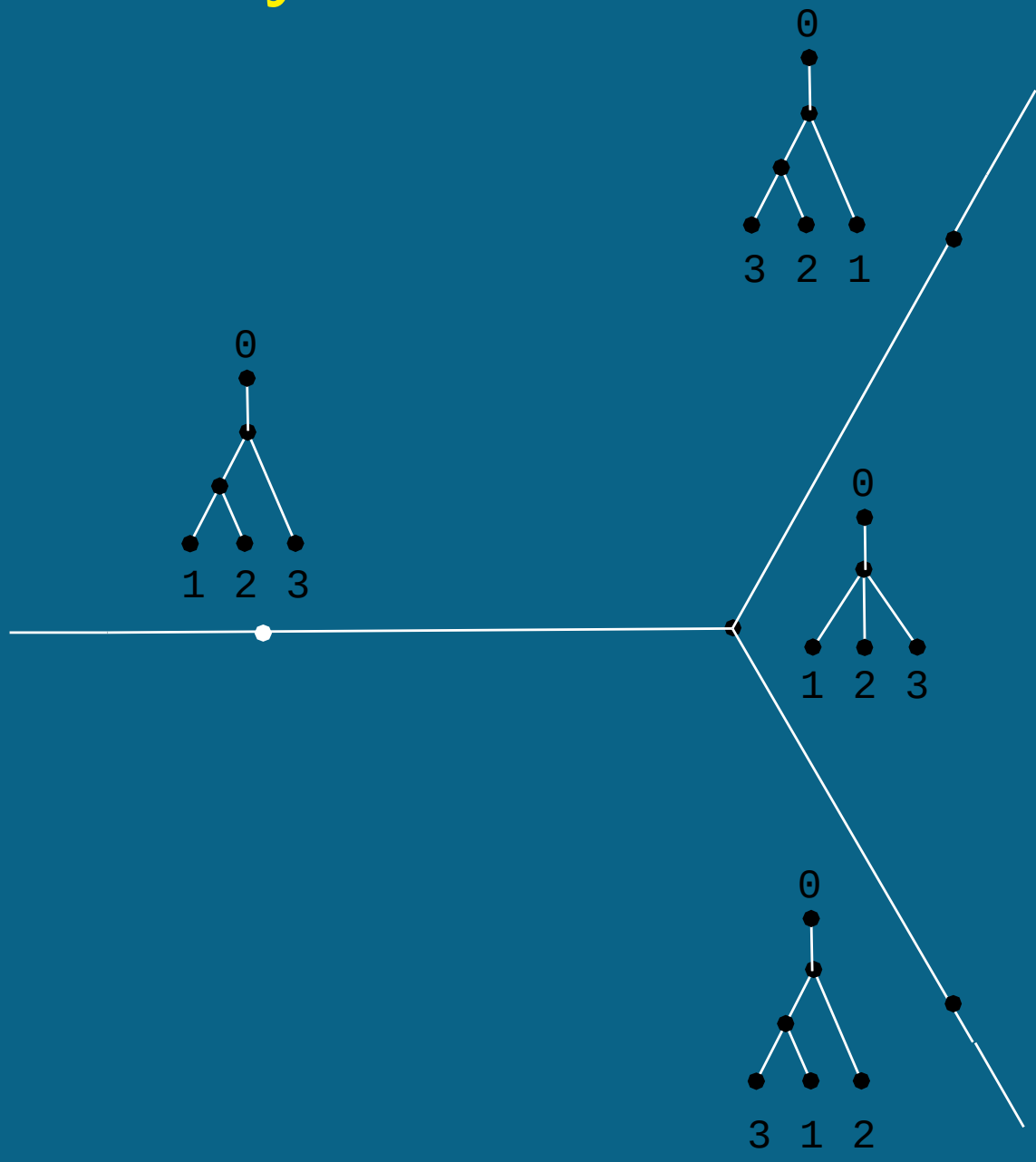


Rotation Moves

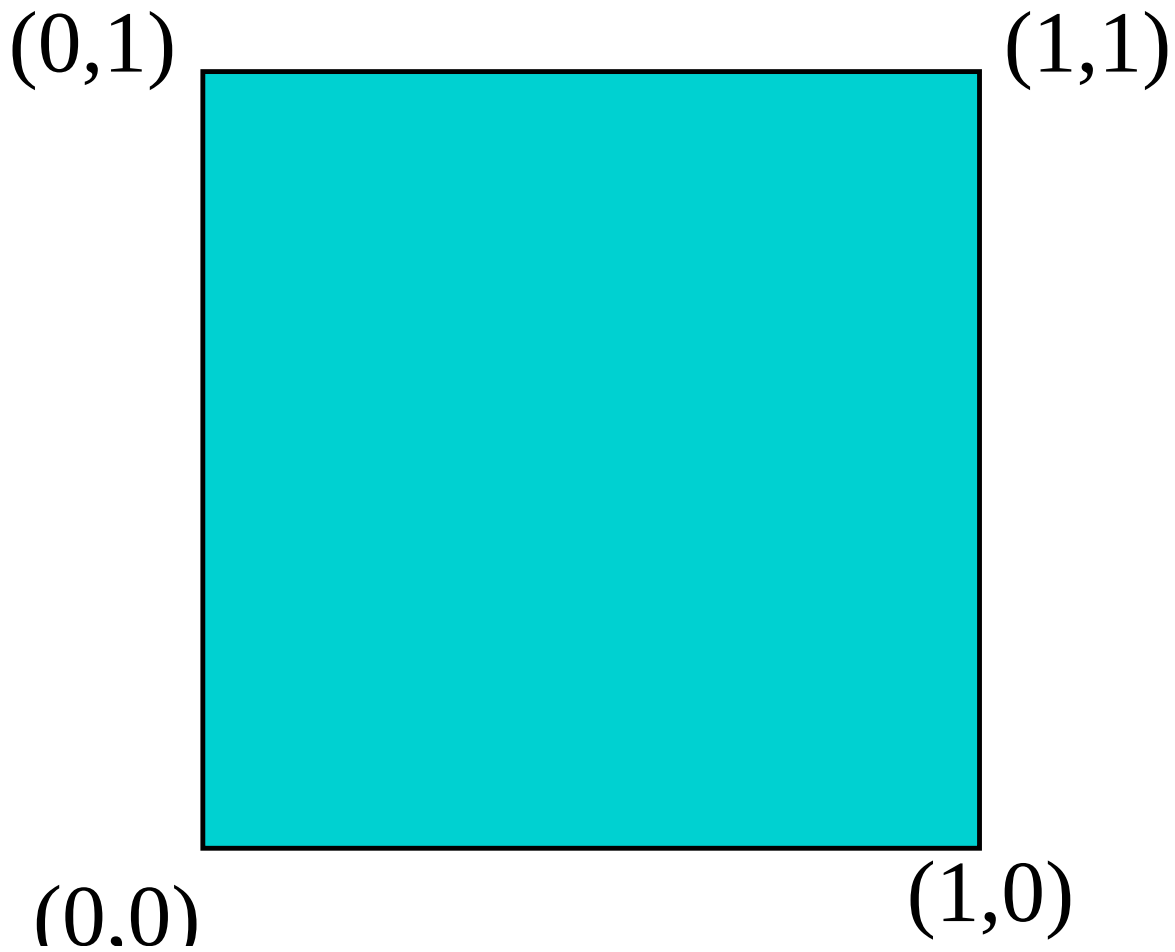


The boundaries between regions represent an area of uncertainty about the exact branching order. In biological terminology this is called an 'unresolved' tree.

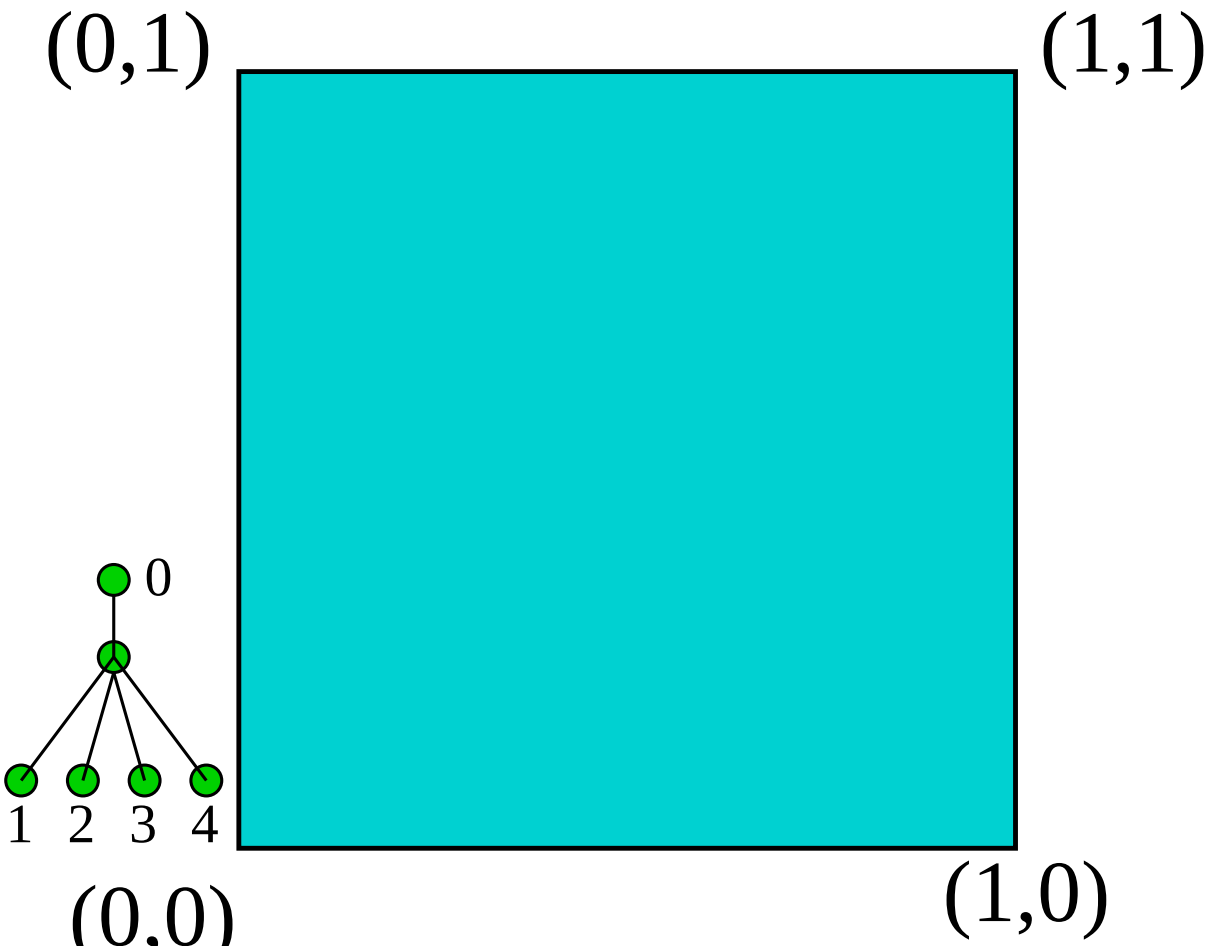
Boundary for trees with 3 leaves



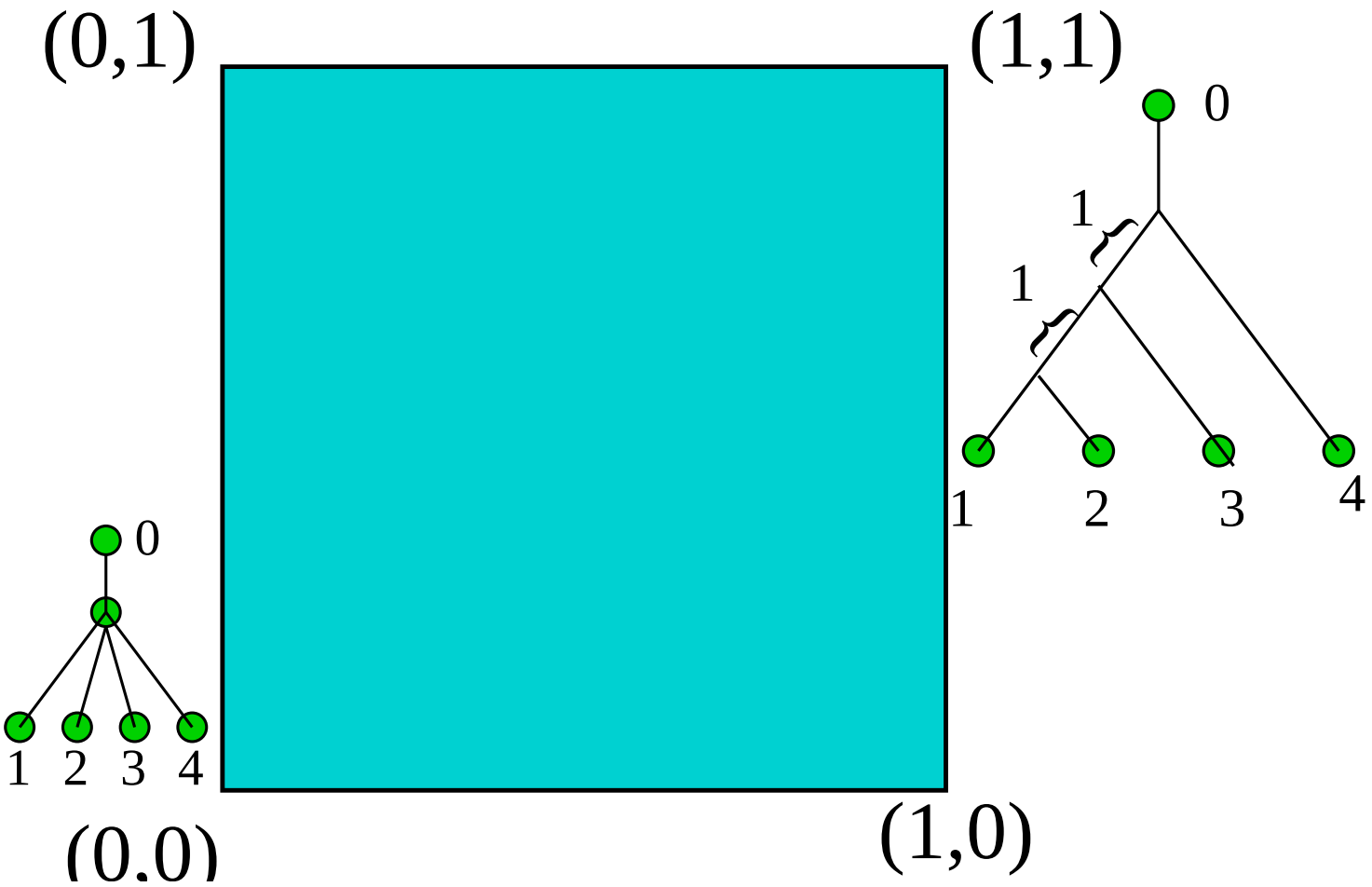
The quadrant for one tree



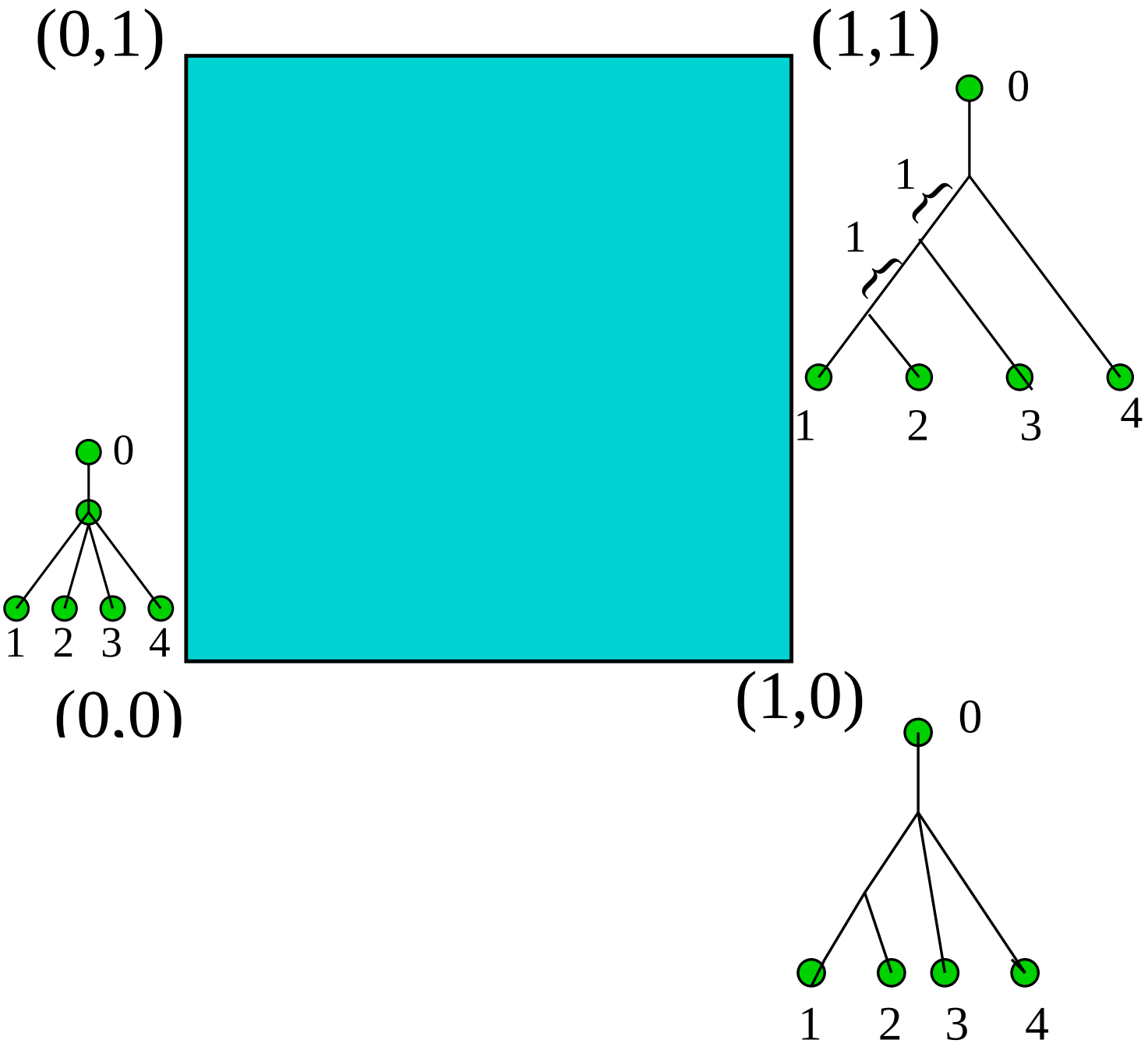
The quadrant for one tree



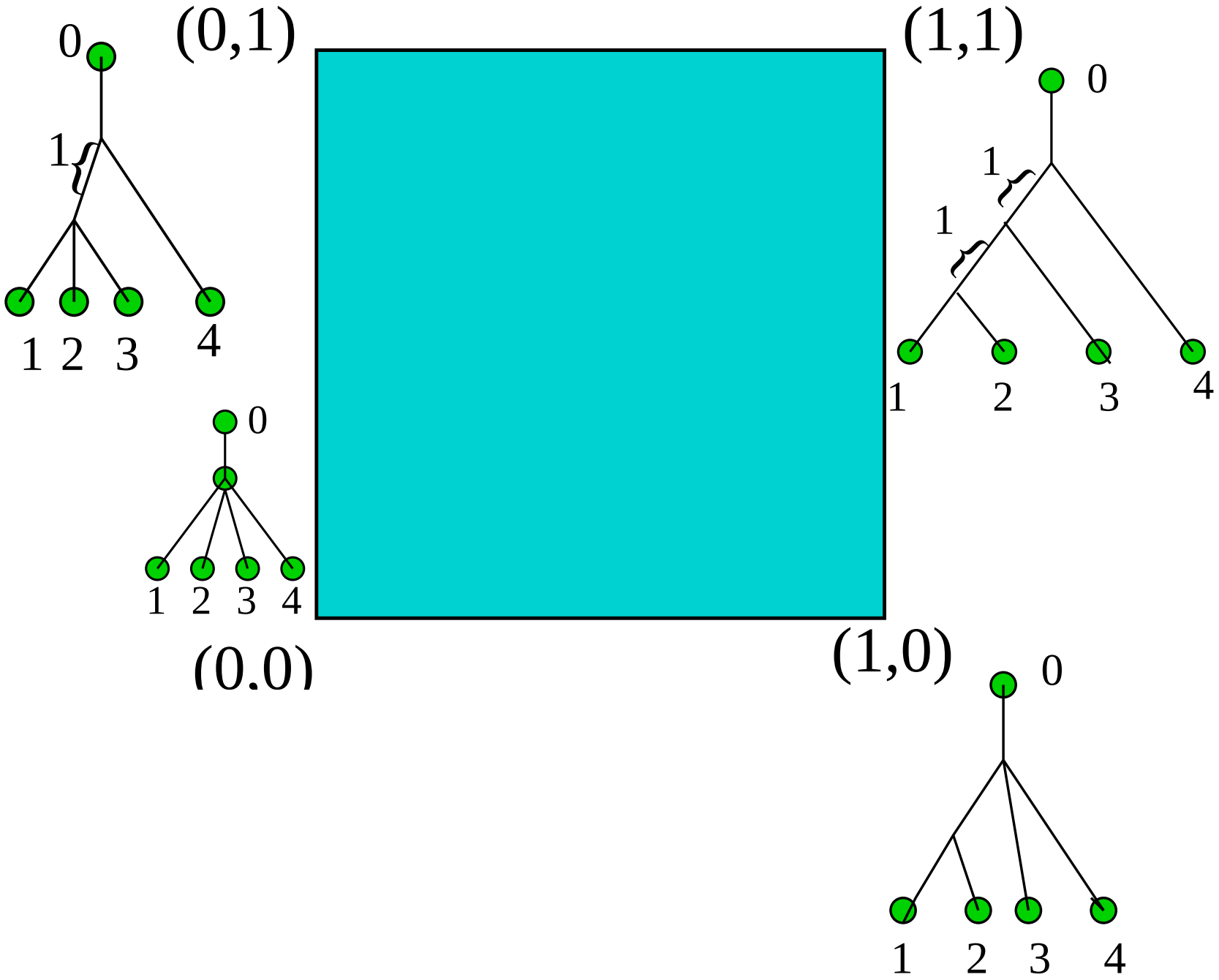
The quadrant for one tree



The quadrant for one tree



The quadrant for one tree



Tree space as a product space

The pendant edges are shared by all trees on the same n leaves, so in fact we decompose treespace into a product space $R^n \times \mathcal{T}_n$.

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

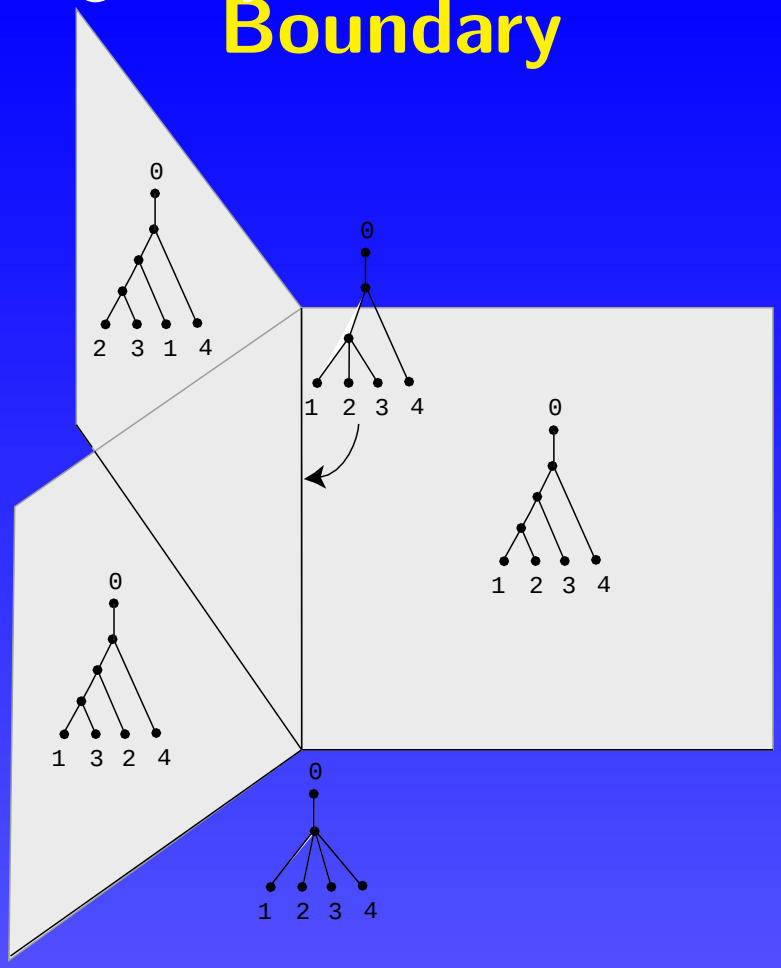
A binary n-tree has the maximal possible number of interior edges $(n-2)$. It determines the largest possible dimensional quadrant which is $n-2$-dimensional.

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.

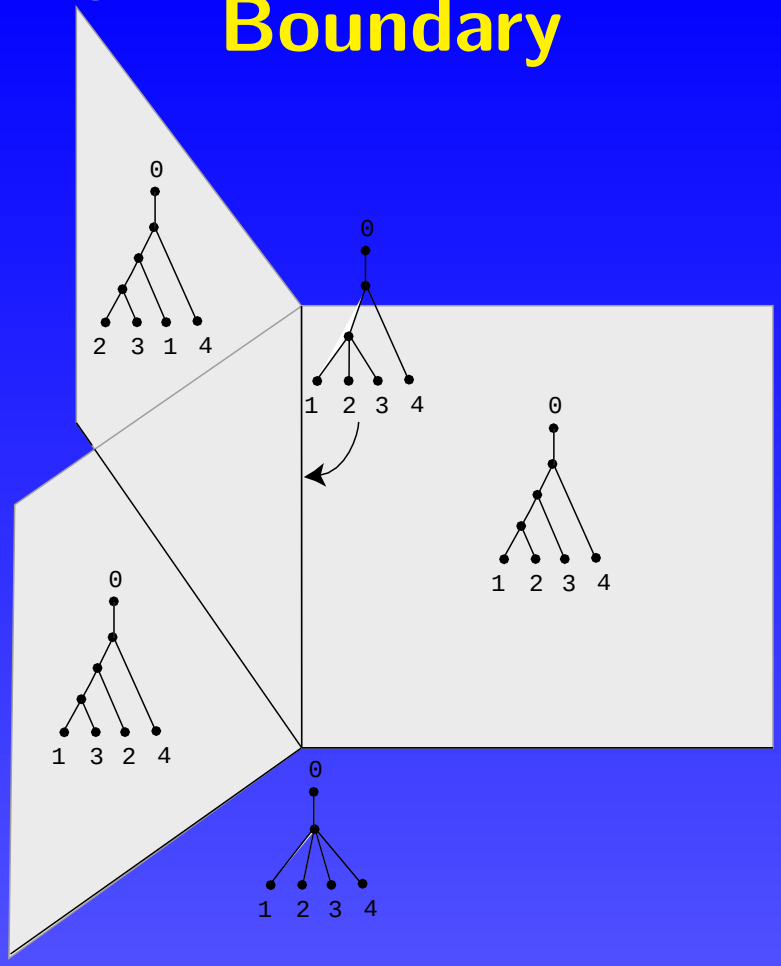
Three quadrants sharing a ray for n=4

Boundary



Three quadrants sharing a ray for n=4

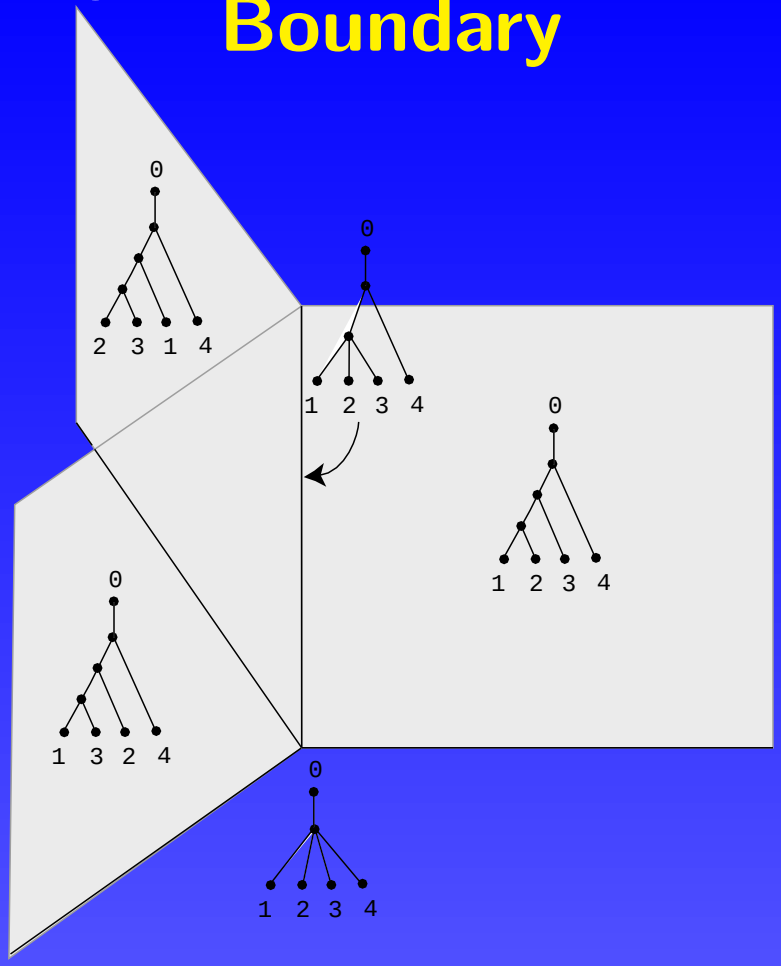
Boundary



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

Three quadrants sharing a ray for n=4

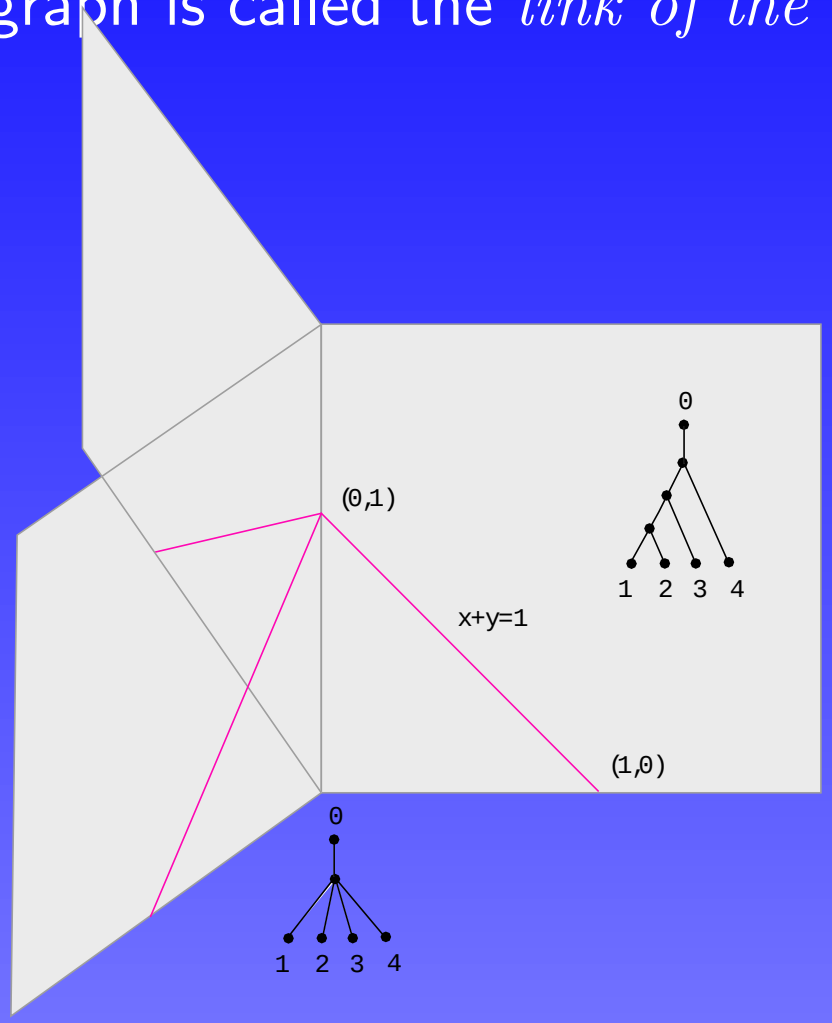
Boundary

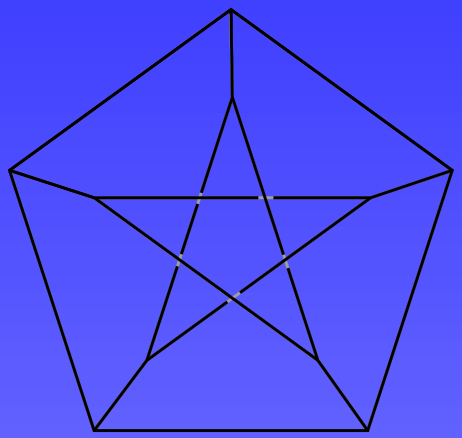


Note that the bottom boundary rays form a copy of $\mathcal{T}_3$ embedded in $\mathcal{T}_4$. In general, $\mathcal{T}_n$ contains many embedded copies of $\mathcal{T}_k$ for $k < n$.

Link of the origin

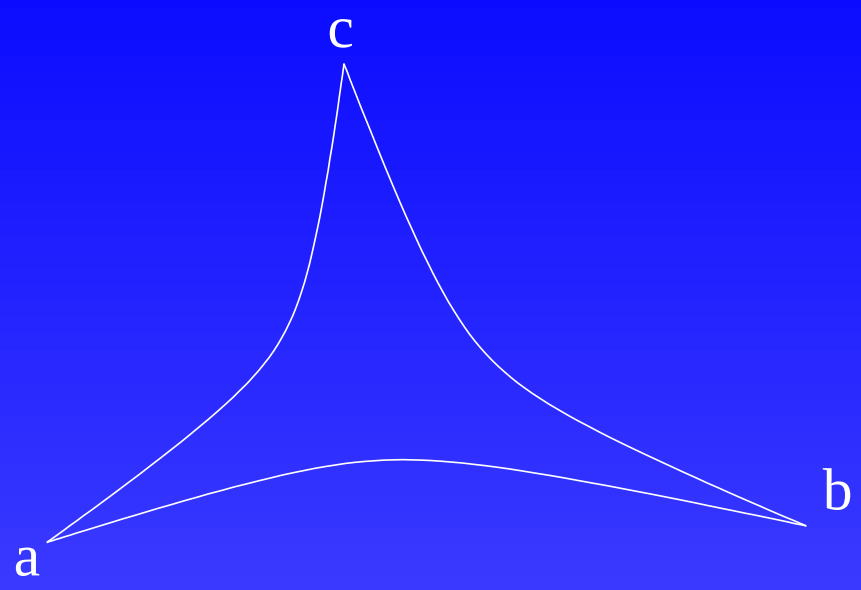
All 15 quadrants for $n = 4$ share the same origin. If we take the diagonal line segment $x + y = 1$ in each quadrant, we obtain a graph with an edge for each quadrant and a trivalent vertex for each boundary ray; this graph is called the *link of the origin*.

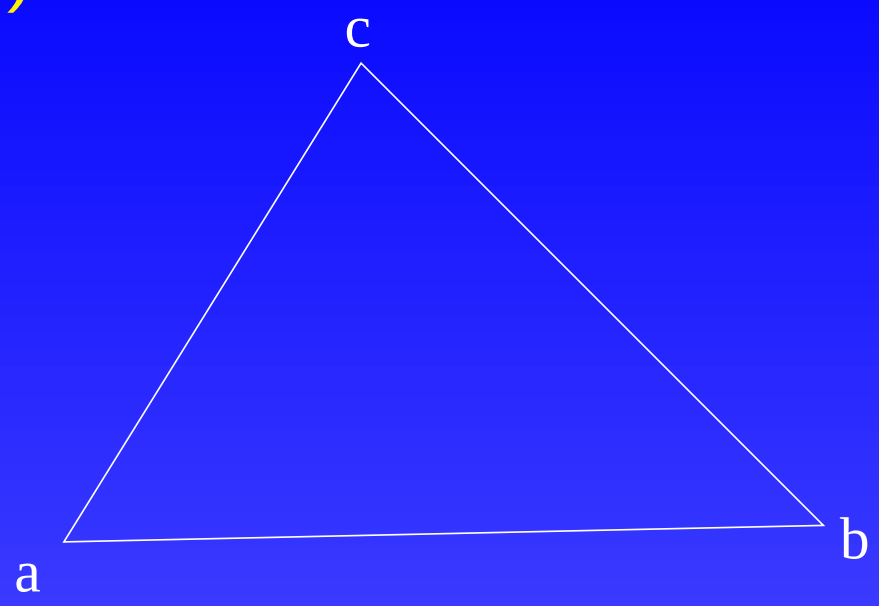




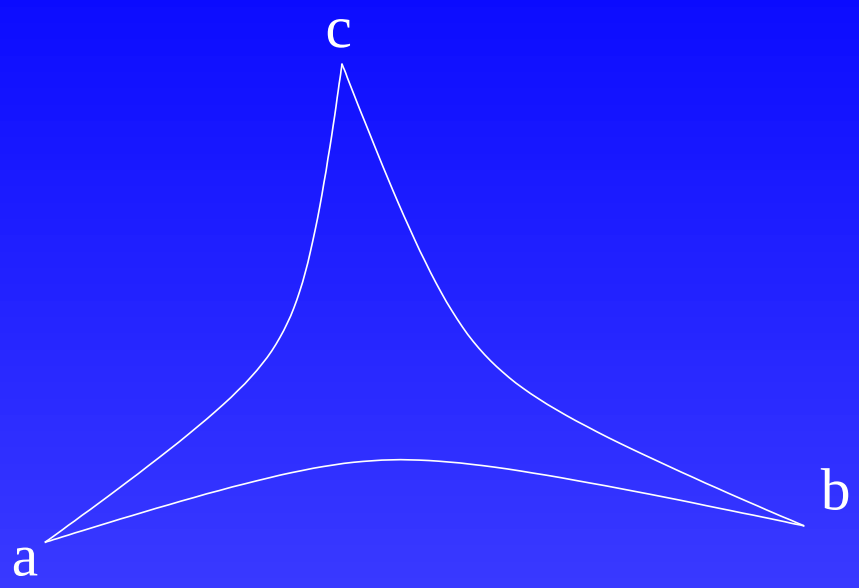
$\longrightarrow$ consequence for time to convergence of MCMC chains of random walks based on NNI moves.

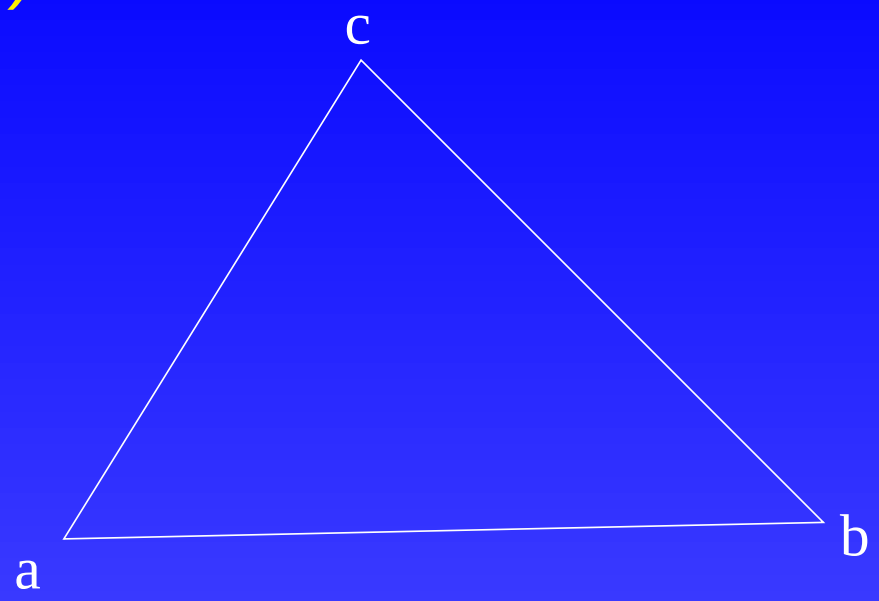
CAT(0) space, (Gromov)





CAT(0) space, (Gromov)





Triangles are thin.

Consequences

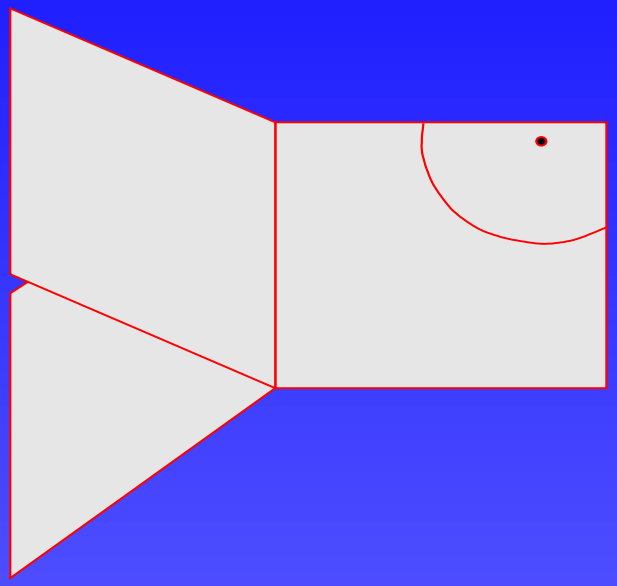
- Averaging works better than it should, (an argument against total evidence computation without decomposing??).
- We can build Bayesian priors based on distances.
- We can make a useful bootstrap statement.
- We can make convex hulls. $\longrightarrow$ Confidence regions.
- We know how many neighbors any tree has.
- We can make a useful bootstrap statement.

How many neighbors for a given tree?(W.H.Li,1993)

We know the number of neighbors of each tree.

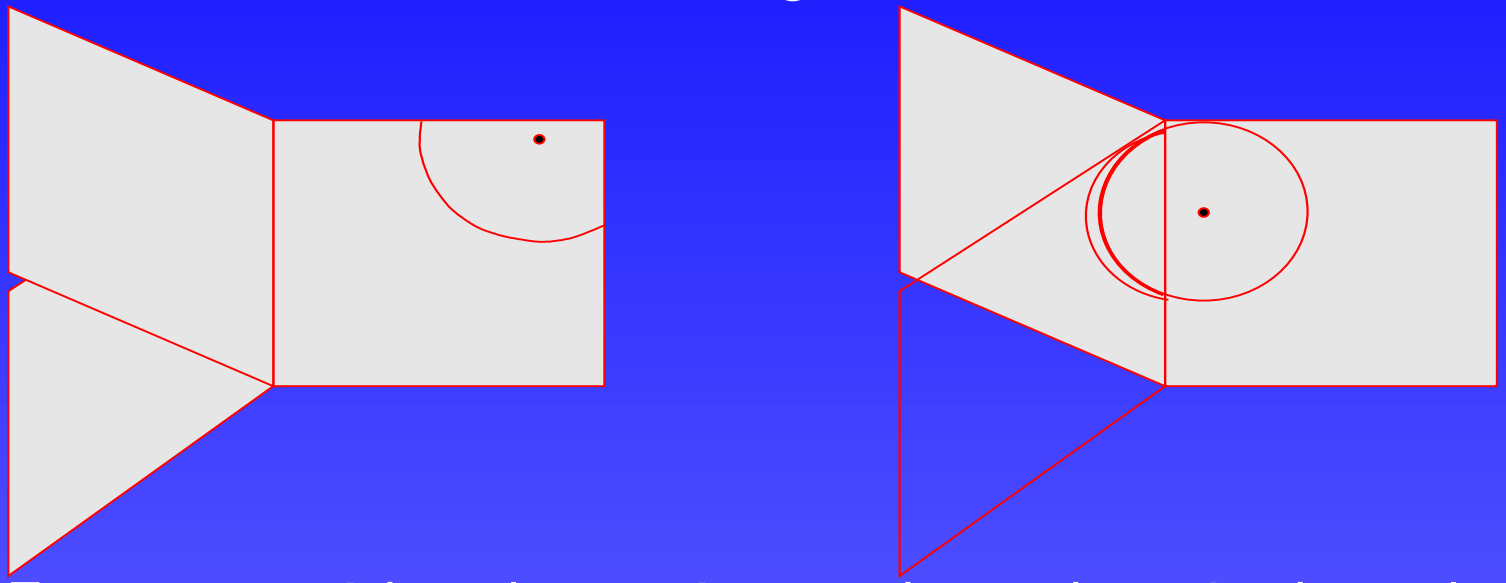
How many neighbors for a given tree?(W.H.Li,1993)

We know the number of neighbors of each tree.



How many neighbors for a given tree?(W.H.Li,1993)

We know the number of neighbors of each tree.

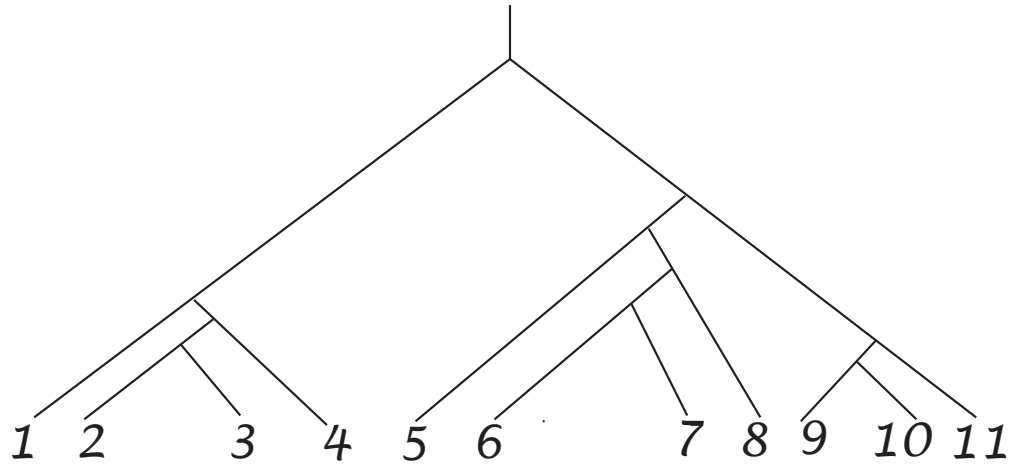


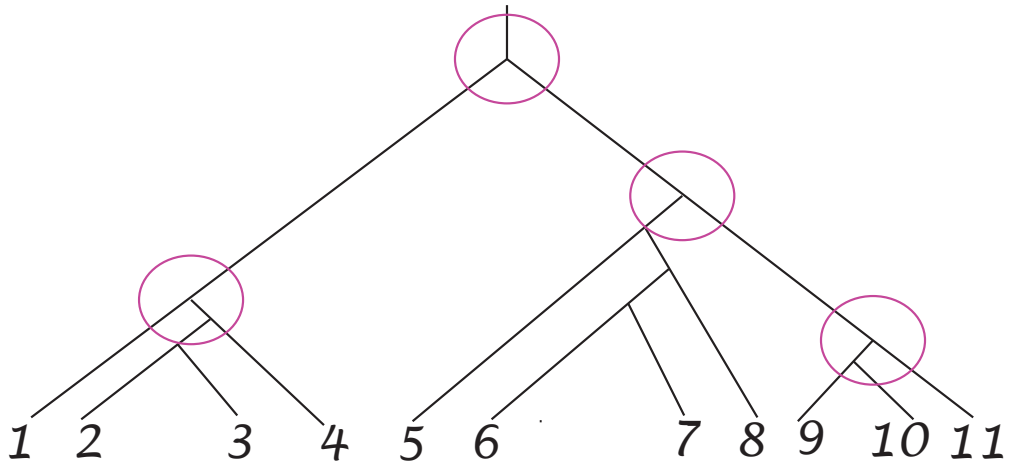
For a tree with only two inner edges, there is the only one way of having two edges small: to be close to the origin-star tree:
15 neighbors. This same notion of neighborhood containing 15 different branching orders applies to all trees on as many leaves as necessary but who have two contiguous “small edges” and all the other inner edges significantly bigger than 0.

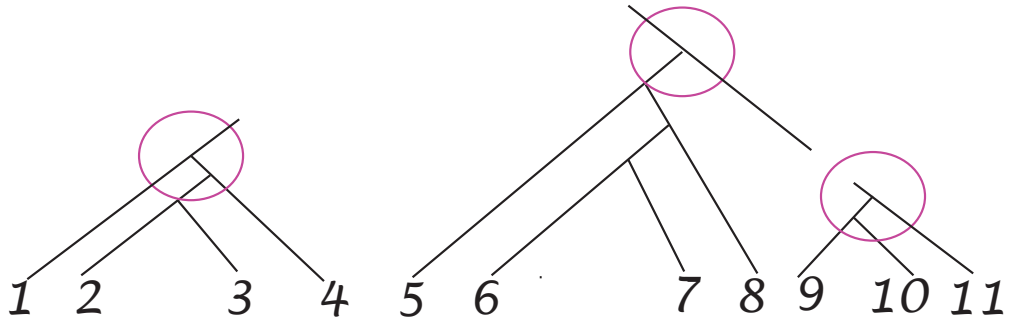
This picture of treespace frees us from having to use simulations to find out how many different trees are in a neighborhood of a given radius r around a given tree. All we have to do is check the sets of continguous edges in the tree smaller than r, say there is only one set of size k, then the neighborhood will contain

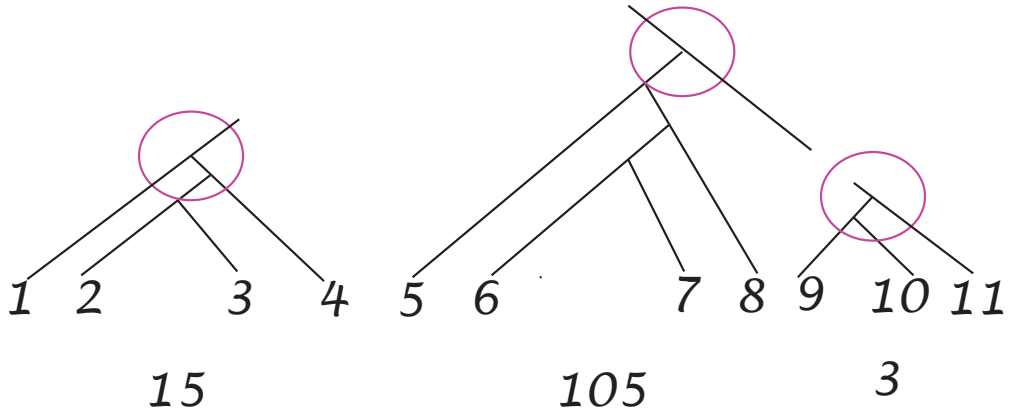
$$(2k-3)!! = (2k-3) \times (2k-5) \times \cdots 3 \text{‘different’ trees.}$$

If there are m sets of sizes $(n_1, n_2, \ldots, n_m)$


1
2
3
4
5
6
7
8
9
10
11


1
2
3
4
5
6
7
8
9
10
11


1
2
3
4
5
6
7
8
9
10
11


1 2 3 4
5 6 7 8 9 10 11
15
105
3

In this case the number of trees within r will be $15 * 105 * 3 = 4725$, in general:

$$(2n_1 - 3)!! \times (2n_2 - 3)!! \times (2n_3 - 3)!! \cdots \times (2n_m - 3)!!$$

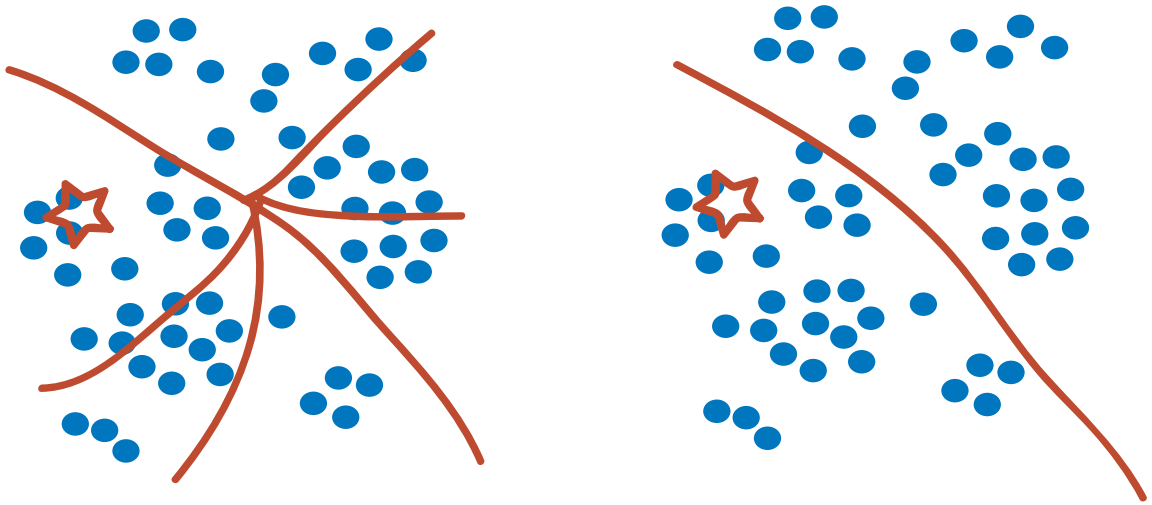
A tree near the star tree at the origin will have an exponential number of neighbors.

This explosion of the volume of a neighborhood at the origin provides for interesting math problems.

These differing number of neighbors for different trees show that the bootstrap values cannot be compared from one tree to another.

This was implicitly understood by Hendy and Penny in their NN Bootstrap procedure.

Are there other ways of using the bootstrap than just counting clade appearances?



Beware the different number of neighbors matters if you think you are using a Monte Carlo method to estimate the distance to the boundary using the bootstrap.

Inferential Bootstrap

$\mathcal{X}$ original data $\longrightarrow \hat{\mathcal{T}}$ estimate.

Call $\mathcal{X}^*$ bootstrap samples consistent with the model used for estimating the tree:

- Non parametric multinomial resampling for a parsimony tree.
- Seqgen parametric type resampling with the same parameters for a ML.
- Bayesian GAMMA prior on rates and generation (Yang 2000) for random sequences according to $\hat{\mathcal{T}}$

Can we use the distance for the bootstrap:

Classical Bootstrap Theory (open problem: to provide such a proof here)

$$\text{Distribution}(d(\hat{\mathcal{T}}, \mathcal{T})) = $$

Can we use the distance for the bootstrap:

Classical Bootstrap Theory (open problem: to provide such a proof here)

$$\text{Distribution}(d(\hat{\mathcal{T}}, \mathcal{T})) = \text{Distribution}\ (d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}}))$$

Can we use the distance for the bootstrap:

Classical Bootstrap Theory (open problem: to provide such a proof here)

$$\text{Distribution}(d(\hat{\mathcal{T}}, \mathcal{T})) = \text{Distribution}\,(d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}}))$$

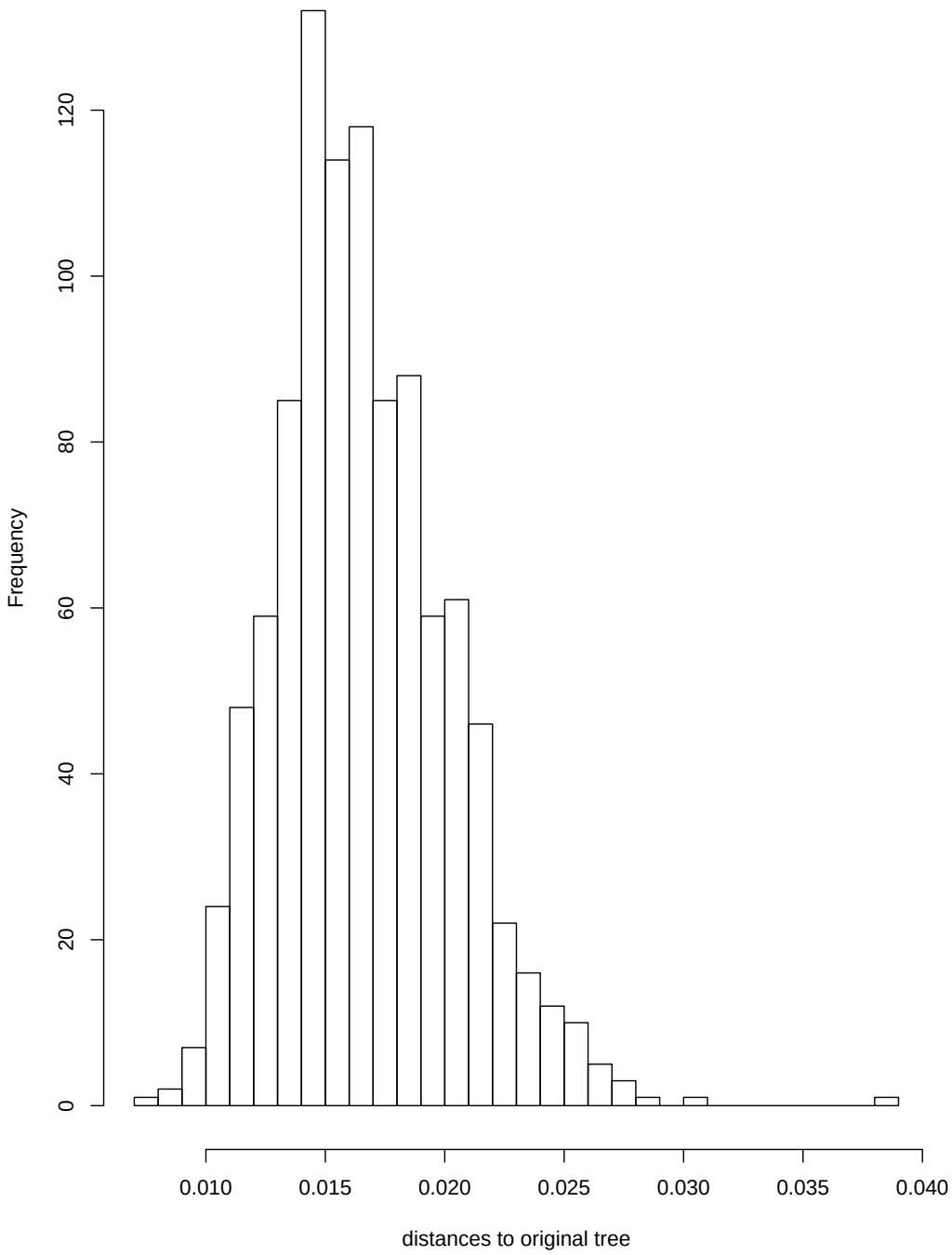
A better pivot than $d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})$?
(classically this would be of the form:

$$\frac{d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})}{sqrt(var(d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})))}$$

What is a variance estimate for trees? $\sum_{i=1}^{n} d(\hat{\mathcal{T}}^*, \hat{\mathcal{T}})^2$?

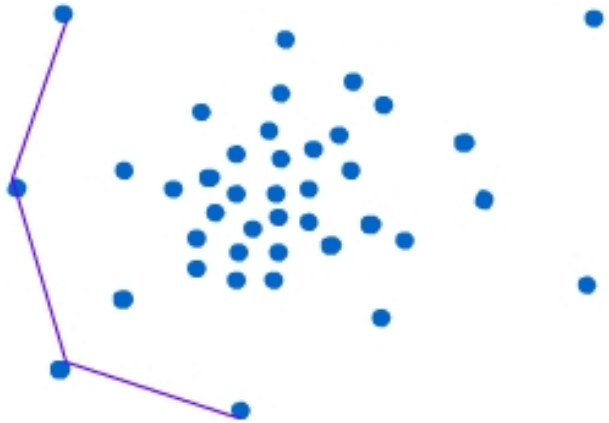
$\longrightarrow$ all open problems

Bootstrap Distribution of Distances to $\hat{\tau}$

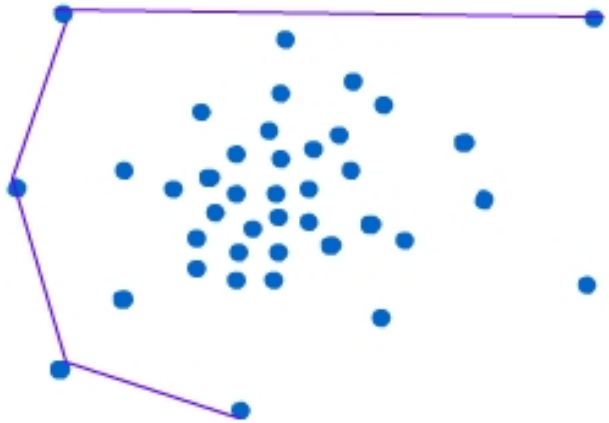

Frequency
0
20
40
60
80
100
120
0.010
0.015
0.020
0.025
0.030
0.035
0.040
distances to original tree

Convex Hulls as Nonparametric Confidence Regions

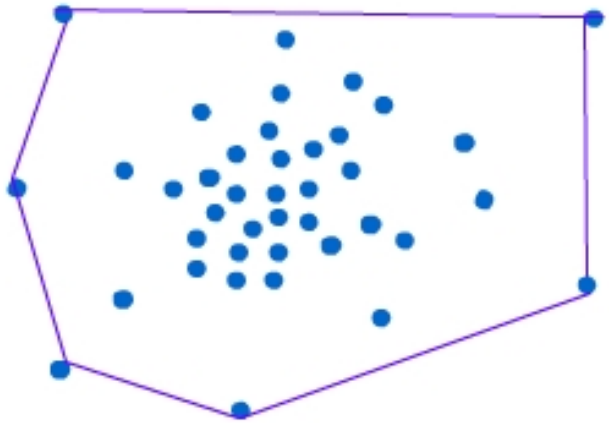
Convex Hulls as Nonparametric Confidence Regions



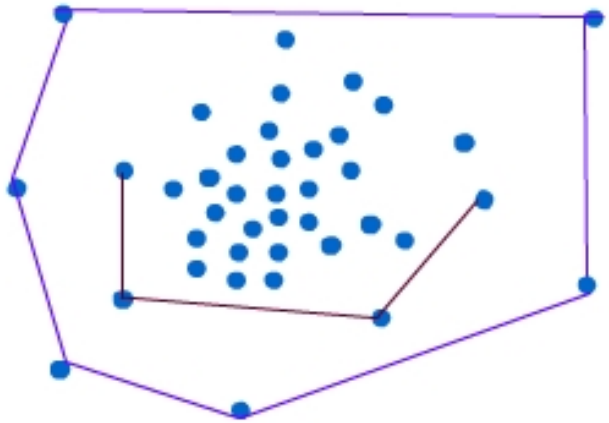
Convex Hulls as Nonparametric Confidence Regions



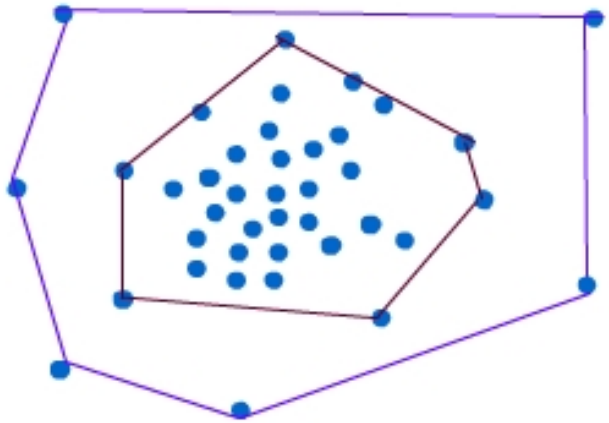
Convex Hulls as Nonparametric Confidence Regions



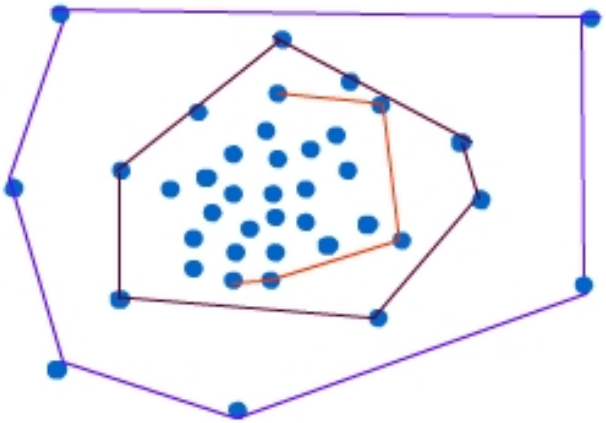
Convex Hulls as Nonparametric Confidence Regions



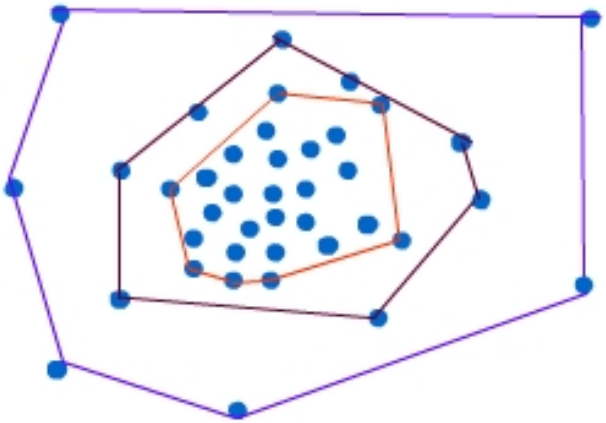
Convex Hulls as Nonparametric Confidence Regions



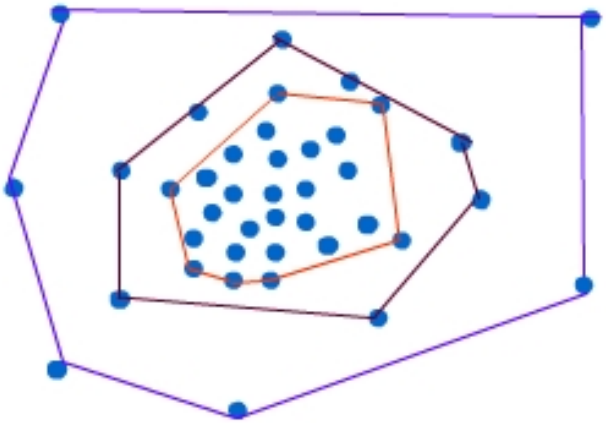
Convex Hulls as Nonparametric Confidence Regions



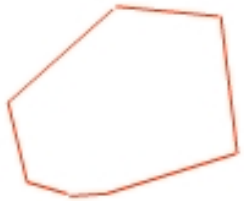
Convex Hulls as Nonparametric Confidence Regions



Convex Hulls as Nonparametric Confidence Regions



Convex Hulls as Nonparametric Confidence Regions



(Tukey's peeling, can also provide a multivariate median if we go all the way down to the inner peel)

There are multivariate versions of this: David Scott's non parametric density estimation shells.

Multivariate Visualizations

Using the distances between trees as input to multidimensional scaling, find a useful 'view' of the trees variability.

Problems: The 'meaning' of the coordinates.

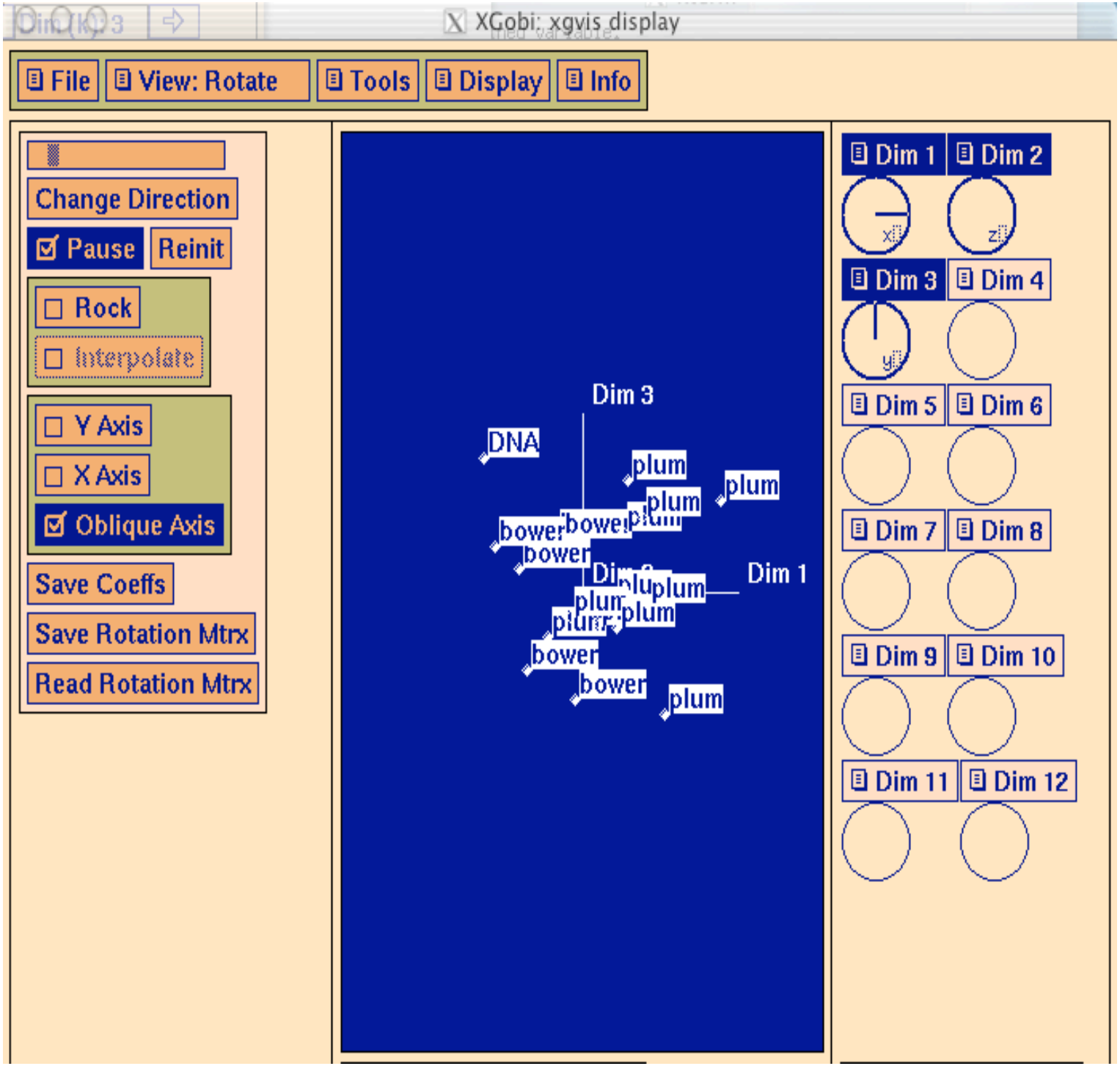
Example: John Endler's bower birds:

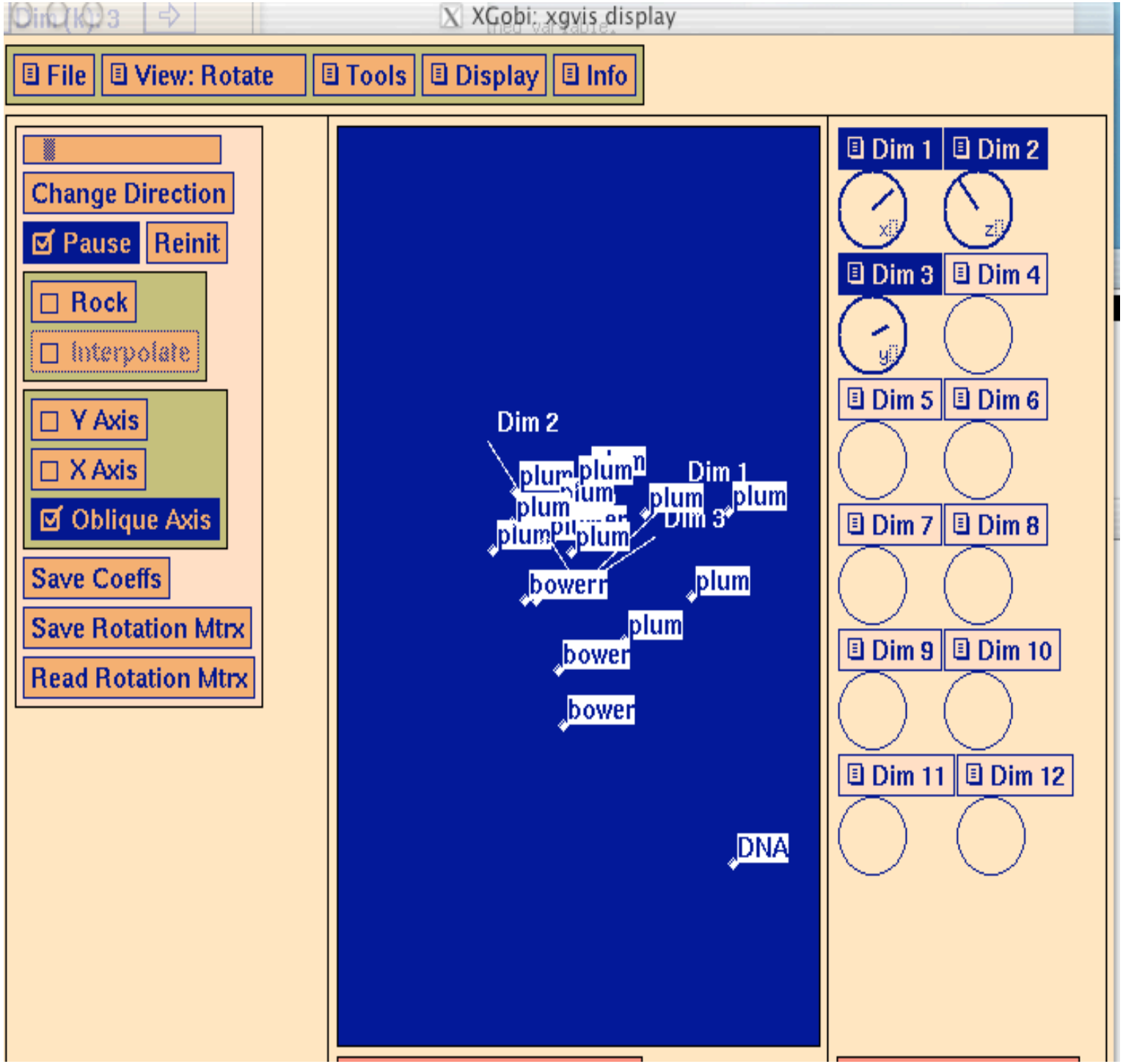
Xgvis Visualisation

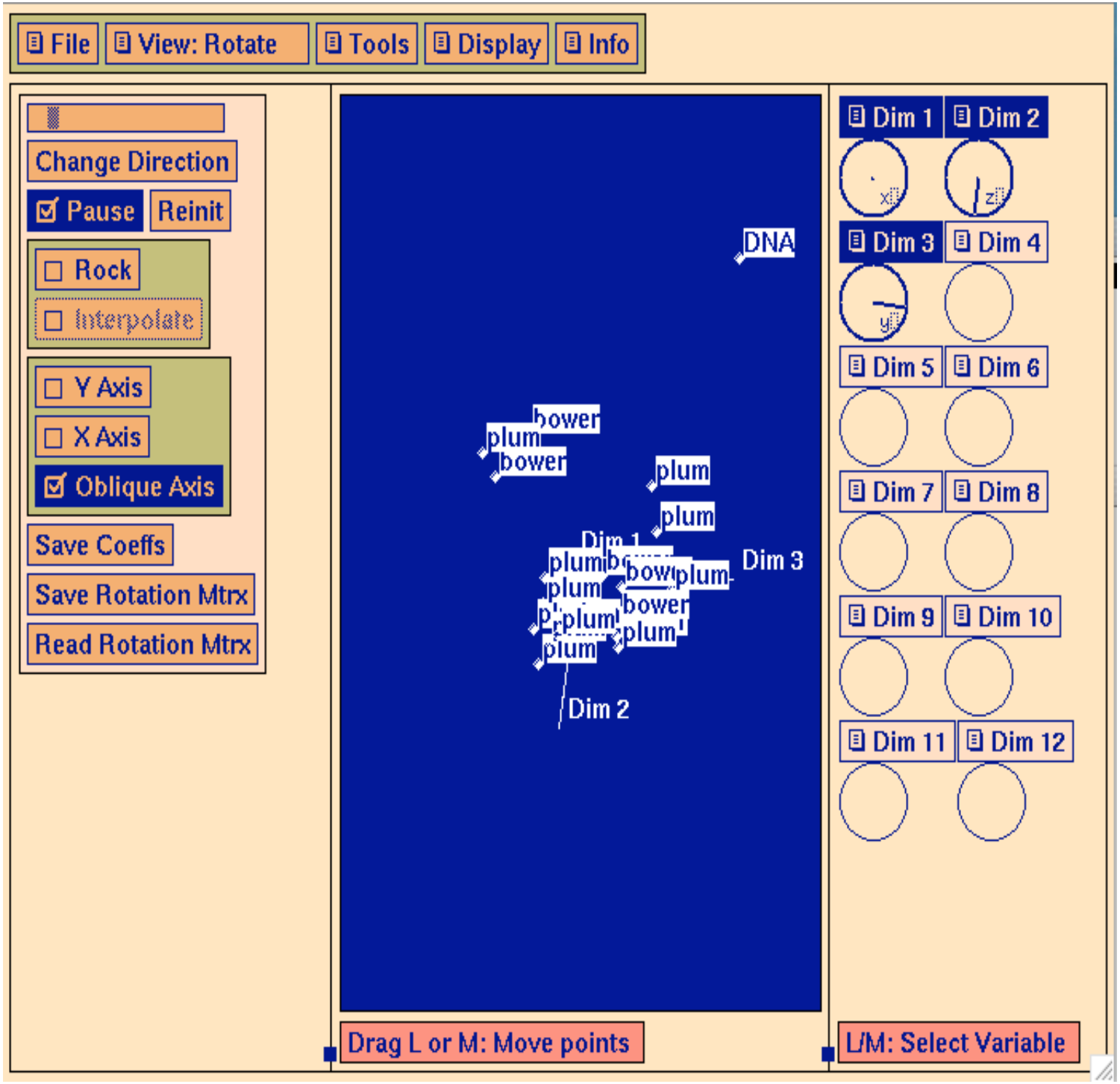
There are many different sources of trees:

- DNA tree.
- Plumage Trees.
- Bower making trees.

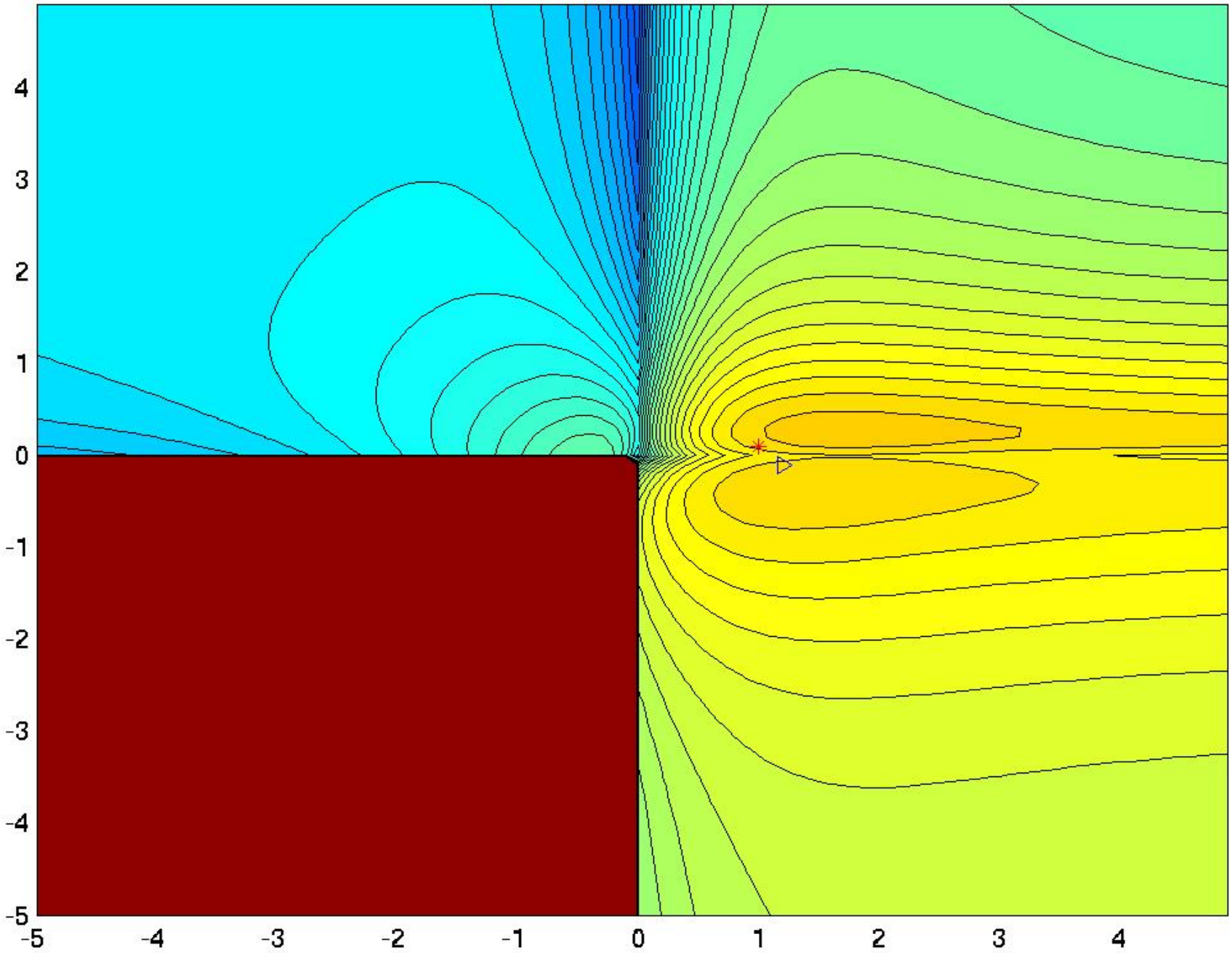
(Interactive plots normally generated by xgvis, here are a few snapshots)


XGobi: xgvis display
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Save Rotation Mtrx
Read Rotation Mtrx
Dim 3
Dim 1
DNA
plum
bower
Dim 1
Dim 2
Dim 3
Dim 4
Dim 5
Dim 6
Dim 7
Dim 8
Dim 9
Dim 10
Dim 11
Dim 12


XGobi: xgvis display
File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Save Rotation Mtrx
Read Rotation Mtrx
Dim 2
Dim 1
Dim 3
plum
bowerr
bower
DNA
Dim 1
Dim 2
Dim 3
Dim 4
Dim 5
Dim 6
Dim 7
Dim 8
Dim 9
Dim 10
Dim 11
Dim 12


File
View: Rotate
Tools
Display
Info
Change Direction
Pause
Reinit
Rock
Interpolate
Y Axis
X Axis
Oblique Axis
Save Coeffs
Save Rotation Mtrx
Read Rotation Mtrx
DNA
bower
plum
bower
plum
plum
Dim 1
Dim 3
Dim 2
Drag L or M: Move points
Dim 1
Dim 2
Dim 3
Dim 4
Dim 5
Dim 6
Dim 7
Dim 8
Dim 9
Dim 10
Dim 11
Dim 12
L/M: Select Variable

Maximum Likelihood Bootstrap


4
3
2
1
0
-1
-2
-3
-4
-5
-5
-4
-3
-2
-1
0
1
2
3
4

Robustness Bootstrap -stability

- Do small perturbations of my data make for changes in the tree?
- How close are the data to being treelike?
- Projection problem (residuals)
- Open problem: we need a notion of differential in treespace to study the influence functions.

How can mathematical statistics help?

Perspectives

- 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)
  - ⋆ Aggregating trees, expectations for various measures.
  - ⋆ Neighborhoods (convex hulls of trees)....
- How much does non-independence matter?
- Justification of commonsense, ground for generalizations.

Answers for David Bryant?

-A tree is rounding the data, if everything else is noise, it purifies the picture, but if there in fact a mixture of 2 trees in the data?

- Data $\longrightarrow$ Tree $+$ Residuals.
Sometimes the residual is all we care about, so losing it is a loss.

- If we have two estimated trees and their sampling distributions $\hat{\tau}_1$ with n leaves $\hat{\tau}_2$ with m leaves, with an intersection of overlapping leaves that is r.

We can include more leaves than actually present by embedding $\mathcal{T}_n$ and $\mathcal{T}_m$ in $\mathcal{T}_k, k > n, m$, all the trees that represents $\hat{\tau_1}$ form a sub-complex of $\mathcal{T}_n$ and we can do the same for $\hat{\tau_2}$, the extra information that we can include is the sampling distributions for both trees. We can try and find an intersection in the support of the two distributions, this is more flexible than just looking for a consensus.

-Spaces of nonpositive curvature are generalisations of hyperbolic spaces (ie trees).

References

[Aldous2001] Aldous, D. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statistical Science* **16**, 23–34.

[Billera et al.2001] Billera, L., Holmes, S., & Vogtmann, K. (2001). The geometry of tree space. *Adv. Appl. Maths* **771–801**.

[Cooper & Penny1997] Cooper, A. & Penny, D. (1997). Mass survival of birds across the cretaceous- tertiary boundary: Molecular evidence. *Science* **275**, 1109–1113.

[Diaconis1989] Diaconis, P. (1989). A generalization of spectral analysis with application to ranked data. *The Annals of Statistics* **17**, 949–979.

[Diaconis & Holmes1998] Diaconis, P. & Holmes, S. (1998). Matchings and phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **95**, 14600–14602 (electronic).

[Efron et al.1996] Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. USA* **93**, 13429–34.

[Green1981] Green, P. J. (1981). Peeling bivariate data. In *Interpreting Multivariate Data*, pages 3– 19.

[Holmes2003] Holmes, S. (2003). Bootstrapping Phylogenetic Trees: Theory and Methods, To appear, Statistical Science.

[Tukey1975] Tukey, J. (1975). Mathematics and the picturing of data. In *Proc. International Congress on Mathmatics*, pages 523–531.

[Yang1994] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from dna sequences with variable rates over sites: approximate methods. *Journal Molecular Evolution* **39**, 306–314.

Proof by direct decomposition

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)[**?**].

$(\mathcal{S}_{2n-2}, \mathcal{B}_{n-1})$ 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 $\lambda = (n-2, 1), S^\lambda$ are the number of times each pair appears. for $\lambda = (n-3, 2), S^\lambda$ are the number of times partition of 4 appears in the tree. for $\lambda = (n-3, 1, 1), S^\lambda$ are the number of times 2 pairs appear simultaneously. This decomposition is similar to what was done by Diaconis for permutation data.[**?**]

References