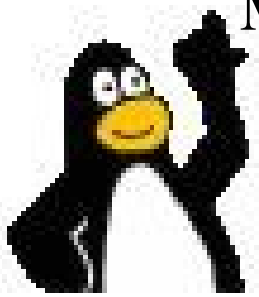
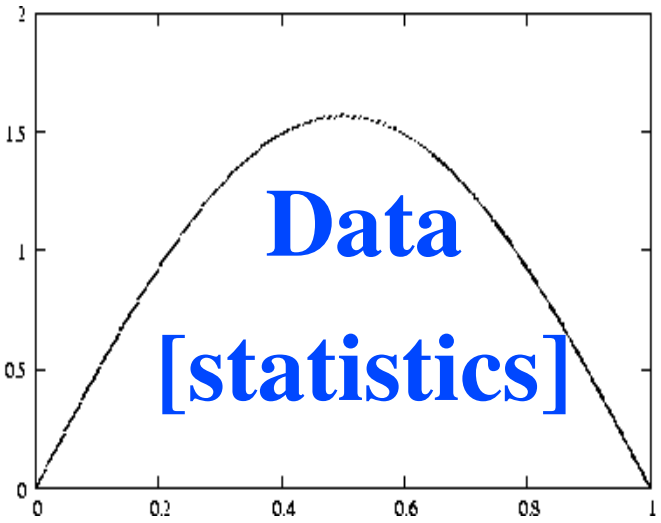
On phylogenetic trees – algebraic geometer's view

Joint project with Weronika Buczyńska
and other students

Microsoft Research Center – Trento, Feb. 16th, 2006

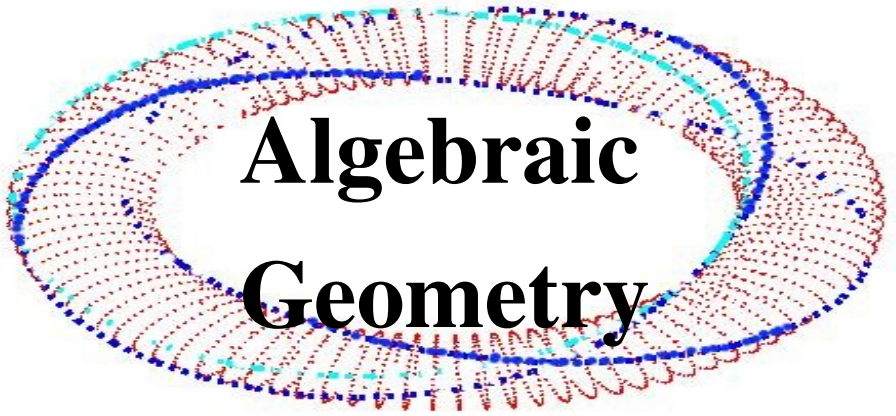


What about algebraic geometry?

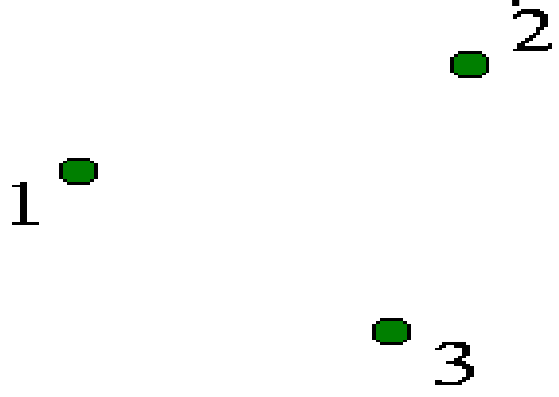


I=ideal(x_3*x_4-x_1*x_7, x_2*x_4-x_1*x_6, x_2*x_3-x_1*x_5, x_2*x_19-x_1*x_31, x_2*x_18-x_1*x_30, x_2*x_17-x_1*x_29, x_2*x_16-x_1*x_28, x_3*x_15-x_1*x_27, x_3*x_14-x_1*x_26, x_3*x_13-x_1*x_25, x_3*x_12-x_1*x_24, x_4*x_11-x_1*x_23, x_4*x_10-x_1*x_22, x_4*x_9-x_1*x_21, x_4*x_8-x_1*x_20, x_2*x_7-x_1*x_32, x_3*x_6-x_1*x_32, x_4*x_5-x_1*x_32, x_6*x_19-x_4*x_31, x_5*x_19-x_3*x_31, x_6*x_18-x_4*x_30, x_5*x_18-x_3*x_30, x_6*x_17-x_4*x_29, x_5*x_17-x_3*x_29, x_6*x_16-x_4*x_28, x_5*x_16-x_3*x_28, x_7*x_15-x_4*x_27, x_5*x_15-x_2*x_27, x_7*x_14-x_4*x_26, x_5*x_14-x_2*x_26, x_7*x_13-x_4*x_25, x_5*x_13-x_2*x_25, x_7*x_12-x_4*x_24, x_5*x_12-x_2*x_24, x_7*x_11-x_4*x_23, x_5*x_11-x_2*x_23, x_7*x_10-x_3*x_22, x_6*x_10-x_2*x_22, x_7*x_9-x_3*x_21, x_6*x_9-x_2*x_21, x_7*x_8-x_3*x_20, x_6*x_8-x_2*x_20, x_6*x_7-x_4*x_32, x_5*x_7-x_3*x_32, x_5*x_6-x_2*x_32, x_17*x_18-x_16*x_19, x_15*x_18-x_14*x_19, x_13*x_18-x_12*x_19, x_11*x_17-x_10*x_19, x_9*x_17-x_8*x_19, x_15*x_16-x_14*x_17, x_13*x_16-x_12*x_17, x_11*x_16-x_10*x_18, x_9*x_16-x_8*x_18, x_13*x_14-x_12*x_15, x_11*x_14-x_10*x_15, x_9*x_14-x_8*x_15, x_11*x_12-x_9*x_14, x_10*x_12-x_8*x_14, x_9*x_11-x_8*x_11, x_19*x_30-x_18*x_31, x_19*x_29-x_17*x_31, x_19*x_28-x_16*x_31, x_7*x_31-x_19*x_32, x_18*x_29-x_16*x_31, x_17*x_30-x_16*x_31, x_18*x_28-x_16*x_30, x_19*x_26-x_18*x_27, x_15*x_30-x_14*x_31, x_19*x_24-x_18*x_25, x_13*x_30-x_12*x_31, x_7*x_30-x_18*x_32, x_17*x_28-x_16*x_29, x_19*x_22-x_17*x_23, x_11*x_29-x_10*x_31, x_19*x_20-x_17*x_21, x_9*x_29-x_8*x_31, x_7*x_29-x_17*x_32, x_17*x_26-x_16*x_27,-x_8*x_14*x_19*x_32, x_1*x_21*x_25*x_31-x_9*x_13*x_19*x_32,

Relations
[algebra]



Example 1: independent variables



Relations in the spaces of triples

$$p_{\alpha\alpha\alpha}p_{\beta\beta\beta} = p_{\alpha\alpha\beta}p_{\beta\beta\alpha} = p_{\alpha\beta\alpha}p_{\beta\alpha\beta} = p_{\beta\alpha\alpha}p_{\alpha\beta\beta}$$

$$p_{\alpha\alpha\alpha}p_{\beta\beta\alpha} = p_{\beta\alpha\alpha}p_{\alpha\beta\alpha}$$

$$p_{\alpha\alpha\alpha}p_{\beta\alpha\beta} = p_{\beta\alpha\alpha}p_{\alpha\alpha\beta}$$

$$p_{\alpha\alpha\alpha}p_{\alpha\beta\beta} = p_{\alpha\beta\alpha}p_{\alpha\alpha\beta}$$

$$p_{\beta\beta\beta}p_{\alpha\alpha\beta} = p_{\alpha\beta\beta}p_{\beta\alpha\beta}$$

$$p_{\beta\beta\beta}p_{\alpha\beta\alpha} = p_{\alpha\beta\beta}p_{\beta\beta\alpha}$$

$$p_{\beta\beta\beta}p_{\beta\alpha\alpha} = p_{\beta\alpha\beta}p_{\beta\beta\alpha}$$

Consider 3 species with complementary features α and β which occur with respective probability $p^i{}_\alpha$ and $p^i{}_\beta$ where $i=1,2,3$. Want to understand the distribution of features α and β in this 3-element system.

If the species are independent then the respective probability is as follows:

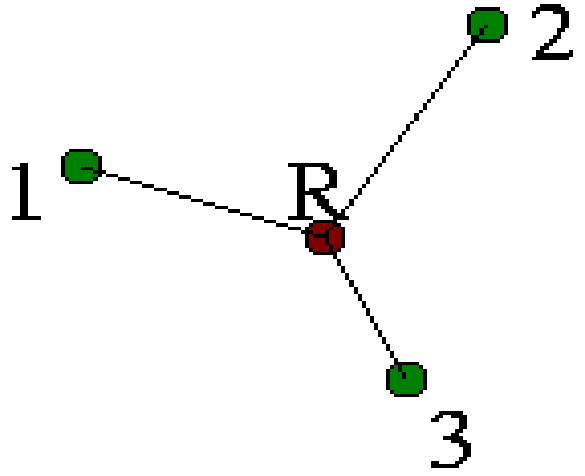
$$p_{\alpha\alpha\alpha} = p^1{}_\alpha p^2{}_\alpha p^3{}_\alpha \qquad p_{\beta\beta\beta} = p^1{}_\beta p^2{}_\beta p^3{}_\beta$$

$$p_{\alpha\alpha\beta} = p^1{}_\alpha p^2{}_\alpha p^3{}_\beta \qquad p_{\beta\beta\alpha} = p^1{}_\beta p^2{}_\beta p^3{}_\alpha$$

$$p_{\alpha\beta\alpha} = p^1{}_\alpha p^2{}_\beta p^3{}_\alpha \qquad p_{\beta\alpha\beta} = p^1{}_\beta p^2{}_\alpha p^3{}_\beta$$

$$p_{\beta\alpha\alpha} = p^1{}_\beta p^2{}_\alpha p^3{}_\alpha \qquad p_{\alpha\beta\beta} = p^1{}_\alpha p^2{}_\beta p^3{}_\beta$$

Example 2: variables in a tree



Now assume that the 3 species had a common ancestor R who had features α and β with probability r_α and r_β (which is not known) and each feature is duplicated in i-th species with probability a_i and changed to the other one with probability b_i

We find out that

$$p_{\alpha\alpha\alpha}=r_\alpha a_1a_2a_3+r_\beta b_1b_2b_3 \quad p_{\beta\beta\beta}=r_\alpha b_1b_2b_3+r_\beta a_1a_2a_3$$

$$p_{\alpha\alpha\beta}=r_\alpha a_1a_2b_3+r_\beta b_1b_2a_3 \quad p_{\beta\beta\alpha}=r_\alpha b_1b_2a_3+r_\beta a_1a_2b_3$$

$$p_{\alpha\beta\alpha}=r_\alpha a_1b_2a_3+r_\beta b_1a_2b_3 \quad p_{\beta\alpha\beta}=r_\alpha b_1a_2b_3+r_\beta a_1b_2a_3$$

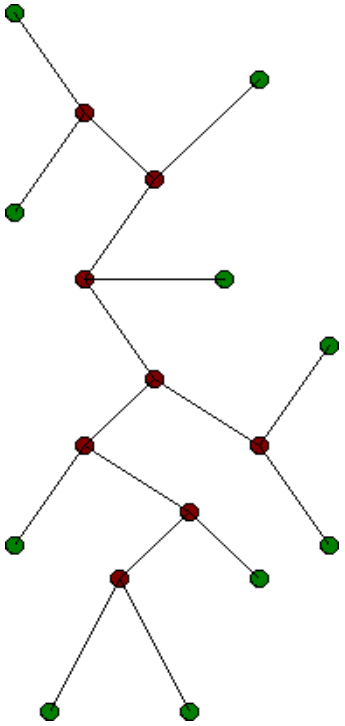
$$p_{\beta\alpha\alpha}=r_\alpha b_1a_2a_3+r_\beta a_1b_2b_3 \quad p_{\alpha\beta\beta}=r_\alpha a_1b_2b_3+r_\beta b_1a_2a_3$$

If $r_\alpha \neq r_\beta$ then via a linear change of coordinates depending on r_α, r_β we arrive to previous equations, otherwise

$$p_{\alpha\alpha\alpha}= p_{\beta\beta\beta} \quad p_{\alpha\alpha\beta}= p_{\beta\beta\alpha}$$

$$p_{\alpha\beta\alpha}= p_{\beta\alpha\beta} \quad p_{\beta\alpha\alpha}= p_{\alpha\beta\beta}$$

Definitions and assumptions



$$\begin{bmatrix} a & b \\ b & a \end{bmatrix}$$

- A tree T is 1-connected graph with the set of edges E and vertices V = *Nodes* + *Leaves*
- At each node we have 3 edges *(3-valent trees)*.
- We consider propagation of two features α and β attached to each vertex of T *(binary trees)*.
- The initial value of probability at a root is uniform and the change of the probability along every edge is symmetric *(symmetric trees)*.
- We observe the distribution of features α and β only on leaves.

Varieties in projective space

Note that all relations in example 1 and 2 are homogeneous and therefore they define a set (variety) in a projective space with coordinates

$$[p_{\alpha\alpha\alpha}, p_{\alpha\alpha\beta}, p_{\alpha\beta\alpha}, p_{\beta\alpha\alpha}, p_{\beta\beta\alpha}, p_{\beta\alpha\beta}, p_{\alpha\beta\beta}, p_{\beta\beta\beta}]$$

Projective space = lines in a vector space. We will consider projective spaces whose coordinates are parametrized by possible distributions of features α and β on all leaves, let us call them the space of parameters.

Projective variety = zero set of homogeneous polynomials in coordinates of the vector space.

Geometric model of a phylogenetic tree

Geometric model of a phylogenetic tree *T*:

projective variety *X*(*T*) in a projective space of parameters which describes geometric locus of probability distributions for a propagation of these features according to the tree *T*.

In general *X*(*T*) is defined by formulas similar to these of Example 2:

$$P\left(X_l = a_{i(l)} : l \in L\right) = \sum_H \left(P\left(X_r = a_{i(r)}\right) \prod_e A^e_{i(s(e)),i(f(e))} \right)$$

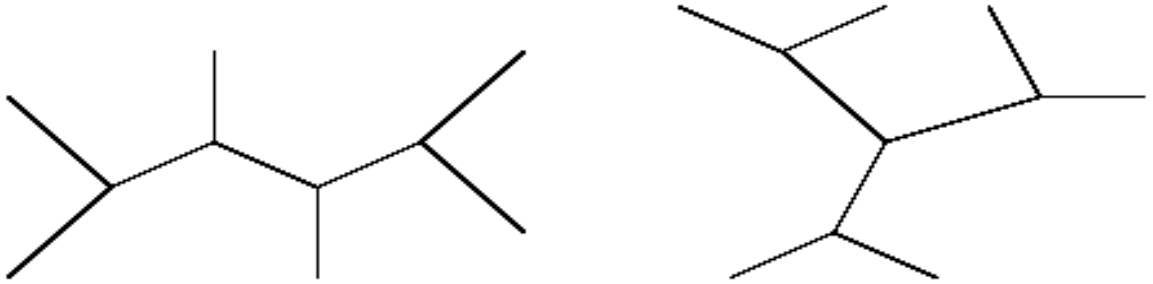
Using algebraic geometry

Understand relation of trees and their geometric models: e.g. find intrinsic invariants of geometric models which distinguish their respective trees.

Example: Hilbert polynomial of a projective variety $X=X(T)$ in $\mathbf{P}^N$ allows (under suitable conditions on X) to predict the number and degree of relations between the distributions of features on leaves.

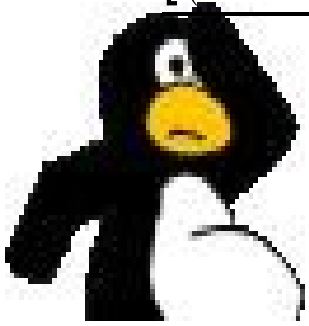
Example: two 3-valent trees with 6 leaves

Can we distinguish these trees by looking at their models?

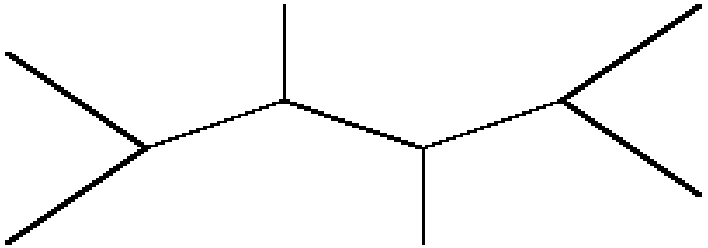


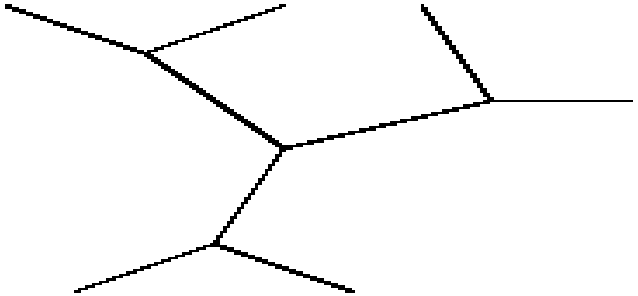
Hilbert polynomial does not distinguish these trees

$$\chi(t) = \frac{1}{22680}(t+1)(t+2)(t+3)(31t^6+374t^5+1942t^4+5616t^3+9511t^2+8988t+3780)$$



Example, continued: are the models distinguishable?





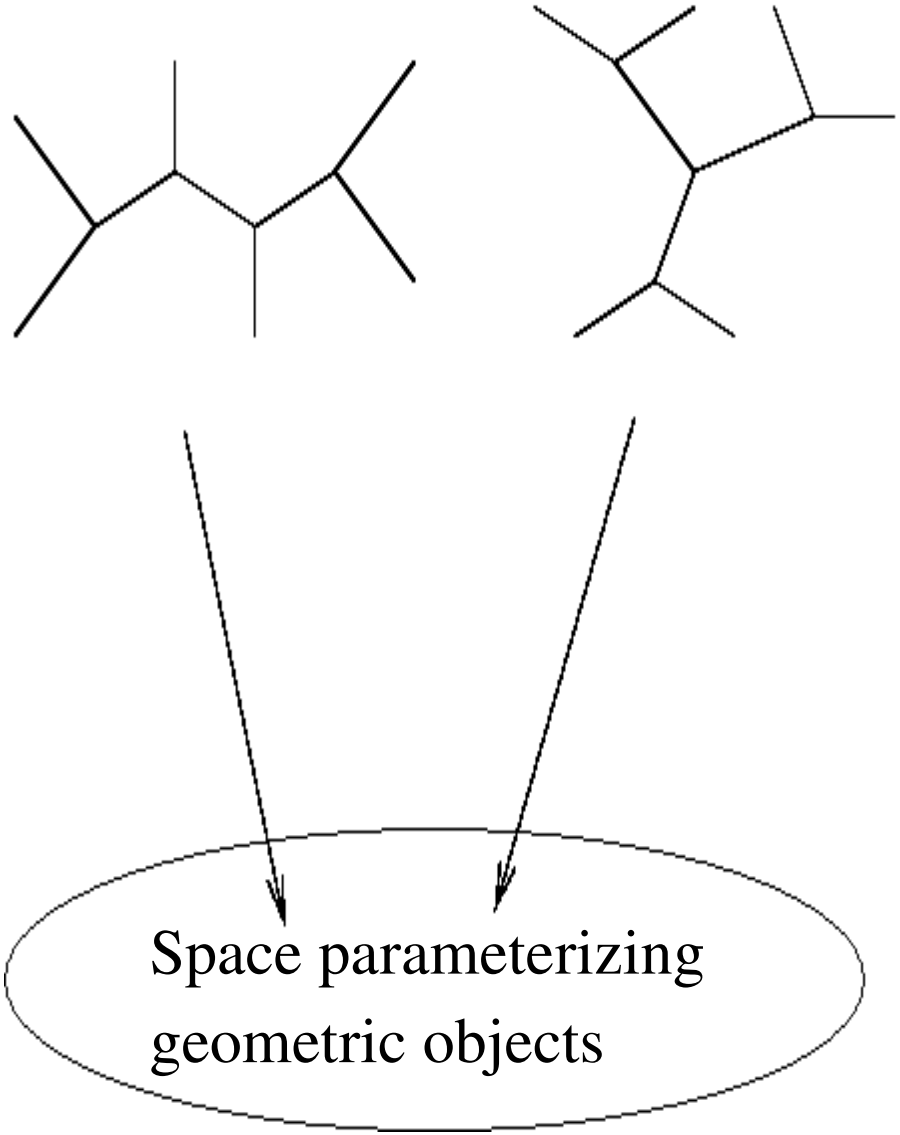
[jarekw@Hermiona jarekw]$ polymake research/drzewa/maximapolymake/3caterpilar.poly
F2_VECTOR
F2_VECTOR
32 480 2400 6144 9312 8832 5280 1920 384
480 240 2400 9456 19904 24896 19104 8816 2240
2400 2400 760 5944 18976 32408 32168 18616 5824
6144 9456 5944 1316 8384 21648 29112 21552 8336
9312 19904 18976 8384 1392 7184 14584 14576 7176
8832 24896 32408 21648 7184 940 3816 5752 3816
5280 19104 32168 29112 14584 3816 406 1224 1224
1920 8816 18616 21552 14576 5752 1224 108 216
384 2240 5824 8336 7176 3816 1224 216 16

[jarekw@Hermiona jarekw]$ polymake research/drzewa/maximapolymake/2star.poly
F2_VECTOR
F2_VECTOR
32 480 2400 6144 9312 8832 5280 1920 384
480 240 2400 9456 19920 24960 19200 8880 2256
2400 2400 760 5944 19008 32552 32408 18792 5872
6144 9456 5944 1316 8400 21744 29308 21720 8388
9312 19920 19008 8400 1392 7200 14640 14640 7200
8832 24960 32552 21744 7200 940 3820 5760 3820
5280 19200 32408 29308 14640 3820 406 1224 1224
1920 8880 18792 21720 14640 5760 1224 108 216
384 2256 5872 8388 7200 3820 1224 216 16

Conclusion: the geometric models are not isomorphic!!

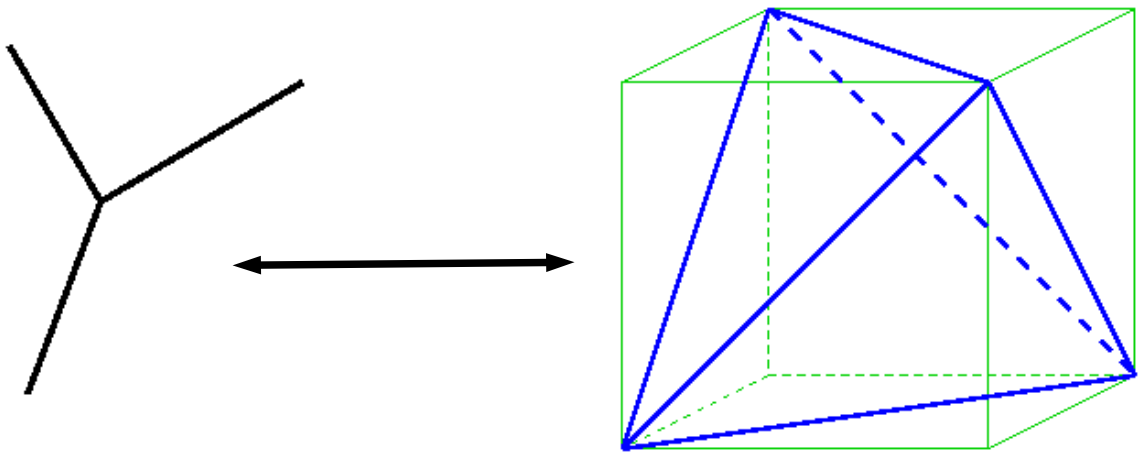
Explanation: deformation.

Theorem. Geometric models of any two binary symmetric 3-valent trees with the same number of leaves are deformation equivalent in the projective space of parameters. In particular equations defining one of them can be deformed to these defining any other one.



Tools: toric geometry

The geometric models of binary symmetric trees are *toric* hence they can be *described* by polytopes with vertices in a lattice.



Two lattices and a polytope

- Given a tree T we consider a free abelian group spanned on the edges of T, that is $M = \oplus_{e \in E} \mathbf{Z}e$
- The vertices of T live in the dual lattice $N=Hom(M,\mathbf{Z})$

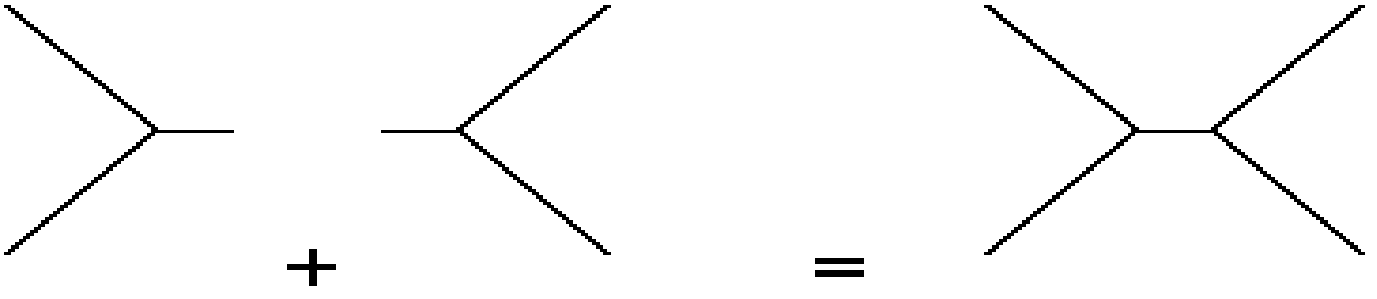
$V=\{v \in M : v(e)=1$ if v is in edge e and $v(e)=0$ otherwise$\}$

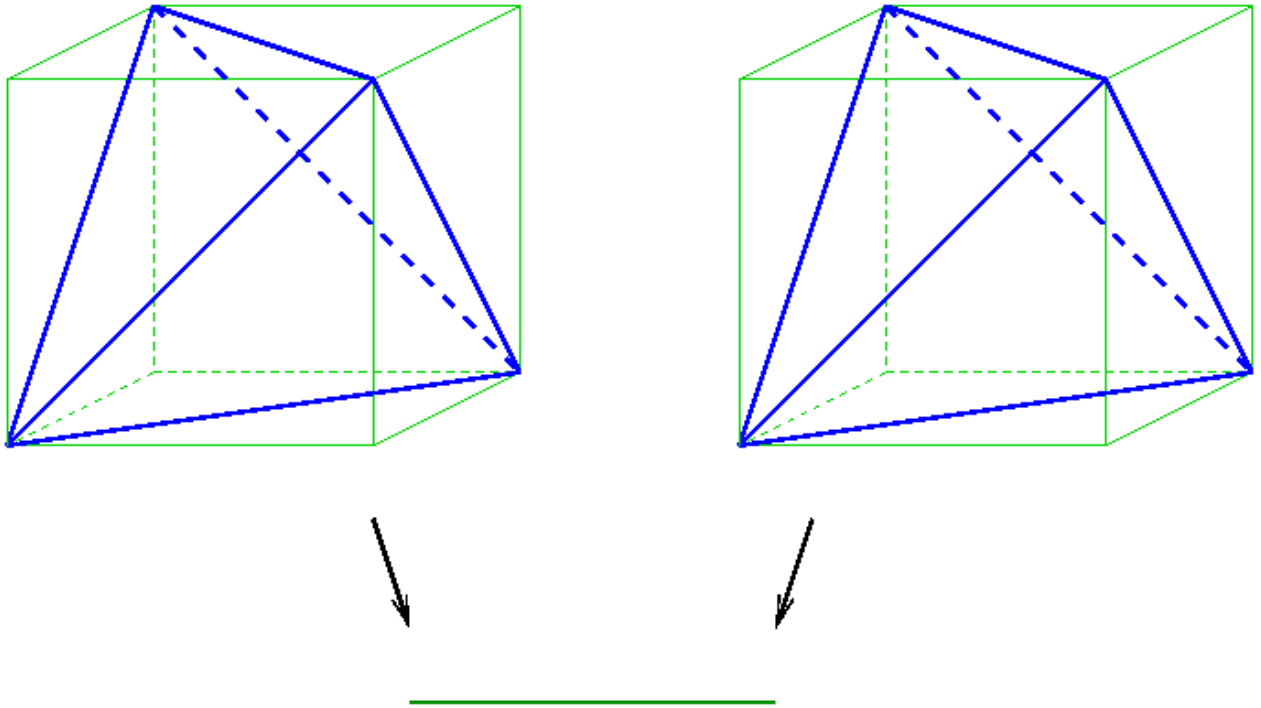
- In the lattice M consider a polytope $\Delta(T)$ with vertices in

$\{u = \Sigma m(e)e : m(e)=0, 1, v(u)=0 \bmod 2$, for all nodes $v\}$

- The polytope $\Delta(T)$ provides a complete description of the variety $X(T)$.

Building up trees: fiber products of polytopes





Conclusion: what about algebraic geometry?

Let $X=X(T)$ in $\mathbf{P}^N$ be a geometric model of a 3-valent binary symmetric tree. Then:

- $(X,O(1))$ is projectively normal
- X has terminal Gorenstein singularities
- X is a Fano variety of index 4, i.e. $K_X=O(-4)$

In dimension 3 any Fano variety with terminal Gorenstein singularities can be smoothened (to a *classical* Fano manifold) [Namikawa]. Does a similar phenomenon hold in higher dimensions?