Combinatorial Distance Geometry: the meeting point between proteins and mathematics

Leo Liberti

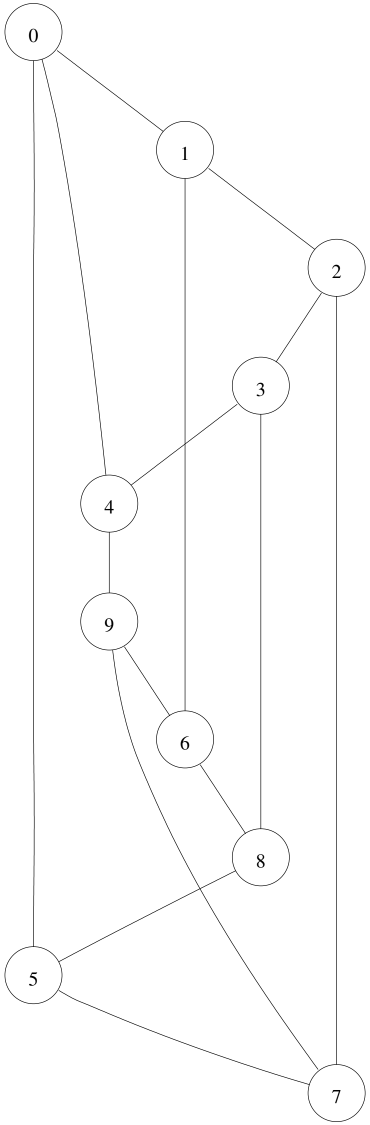
LIX, École Polytechnique, France

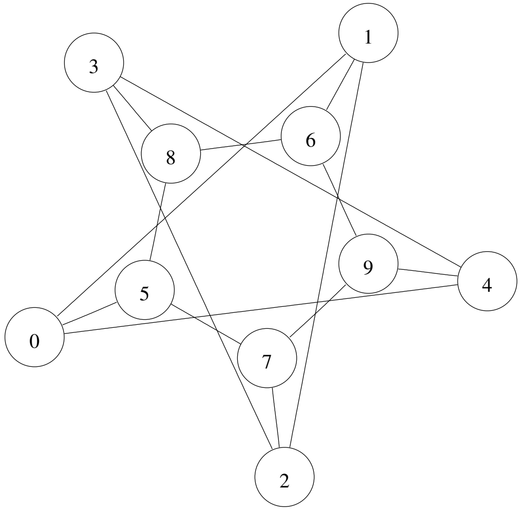
Joint work with:

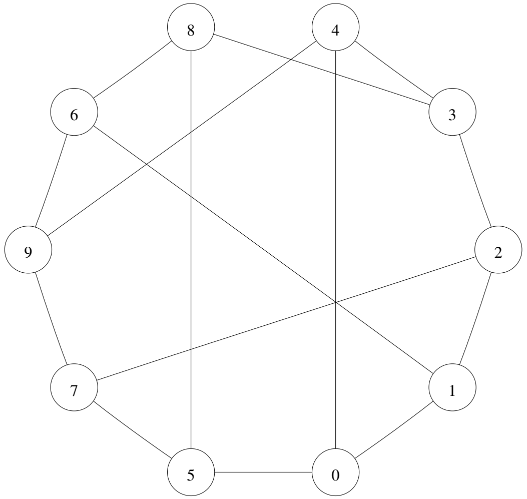
C. Lavor (IMECC-UNICAMP), N. Maculan (COPPE-UFRJ), A. Mucherino (Univ. Rennes)

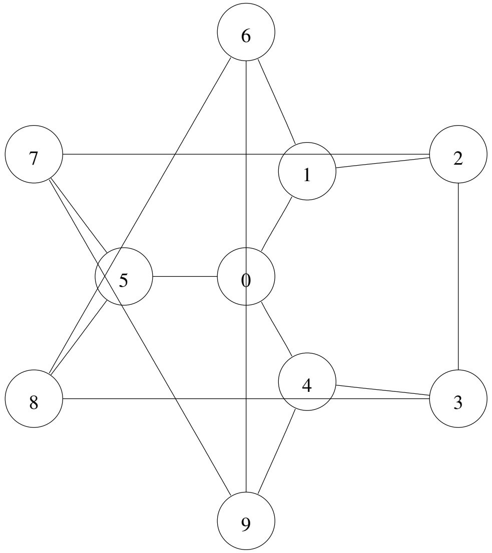
J. Lee (Univ. Michigan), B. Masson (INRIA), M. Nilges (Inst. Pasteur), T. Malliavin (Inst. Pasteur)

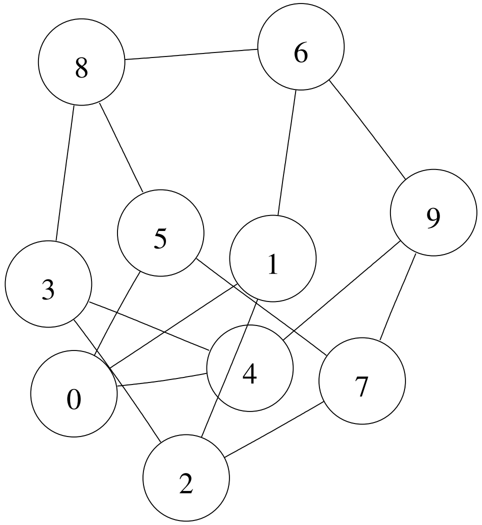
At a glance





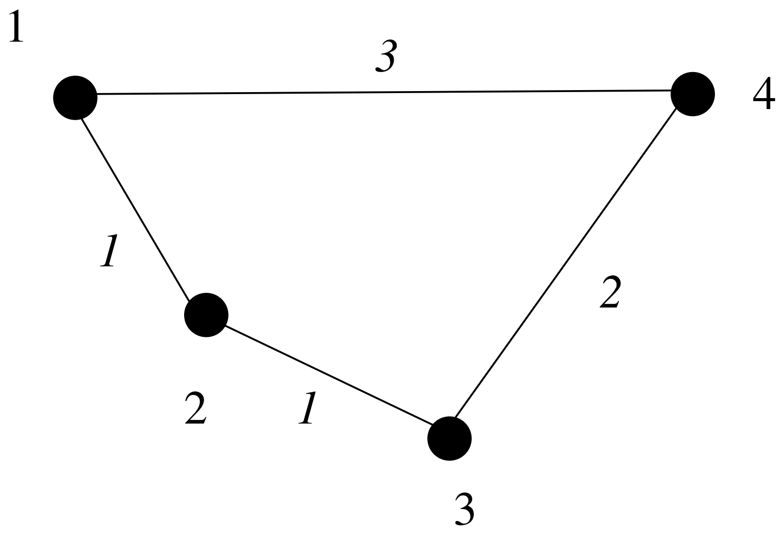






Which graph has most symmetries?

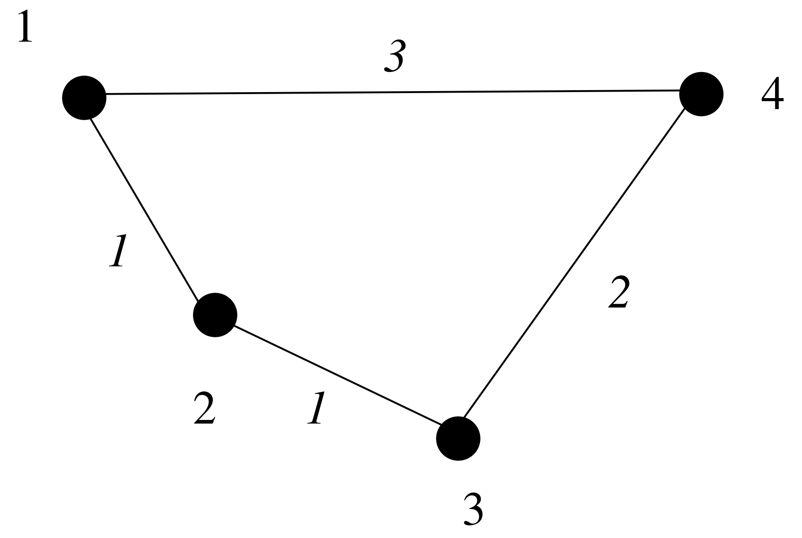
How does a weighted graph look?



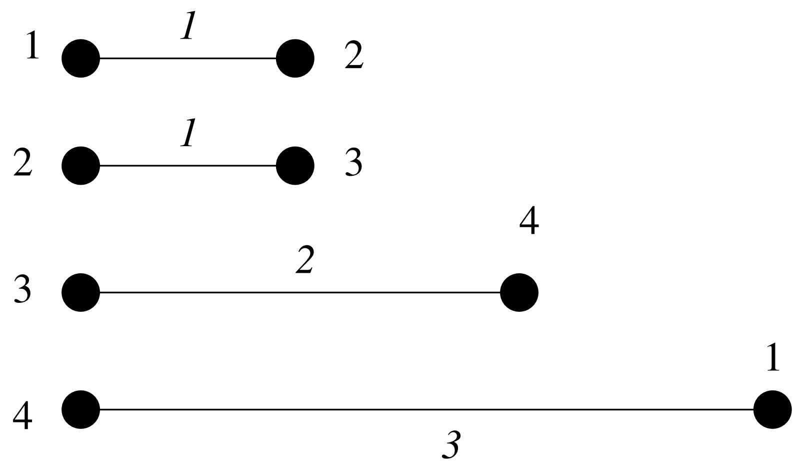
- Like this?

How does a weighted graph look?

- Like this?



- Perhaps like this?



Drawing a graph

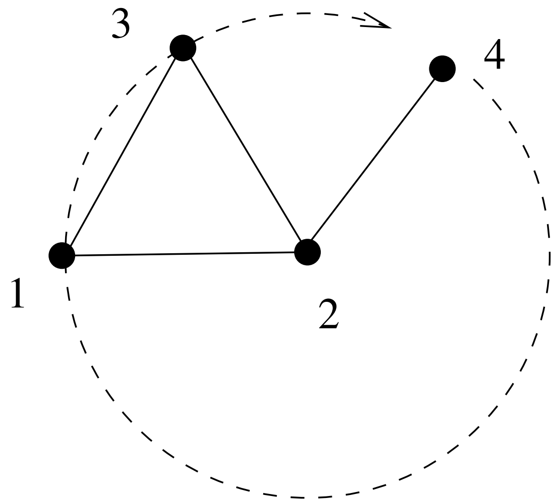
- Given a simple weighted undirected graph $G = (V, E)$ with a distance function $d : E \to \mathbb{R}_+$, solve the constraint system:

$$\forall \{u, v\} \in E \quad \|x_u - x_v\| = d_{uv} \tag{1}$$

- Obtain an embedding $x : V \to \mathbb{R}^2$

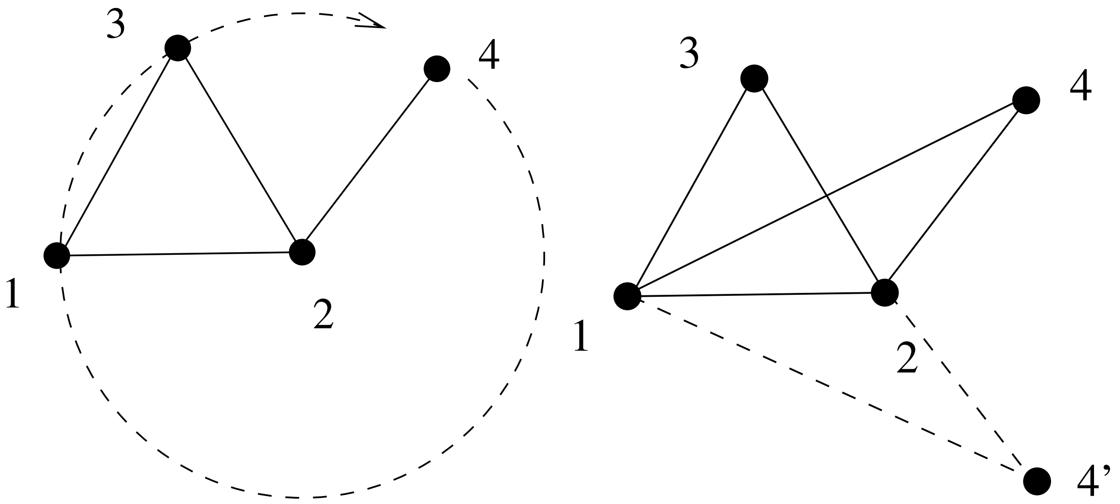
The number of embeddings

- Certain graphs have uncountably many (incongruent) embeddings



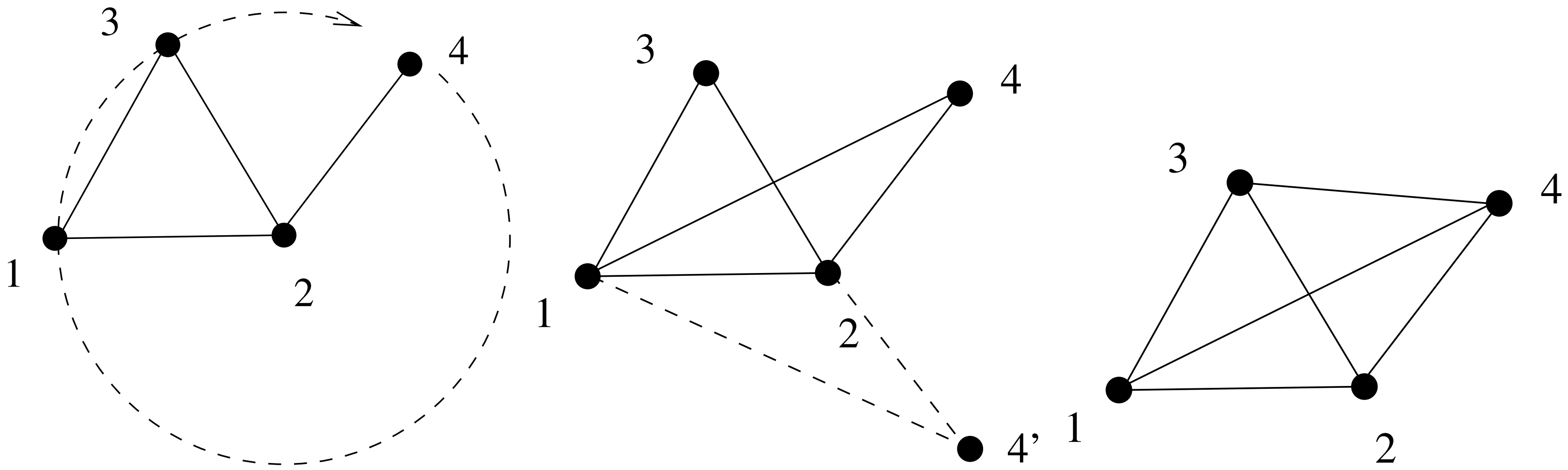
The number of embeddings

- Certain graphs have uncountably many (incongruent) embeddings
- Others have finitely many



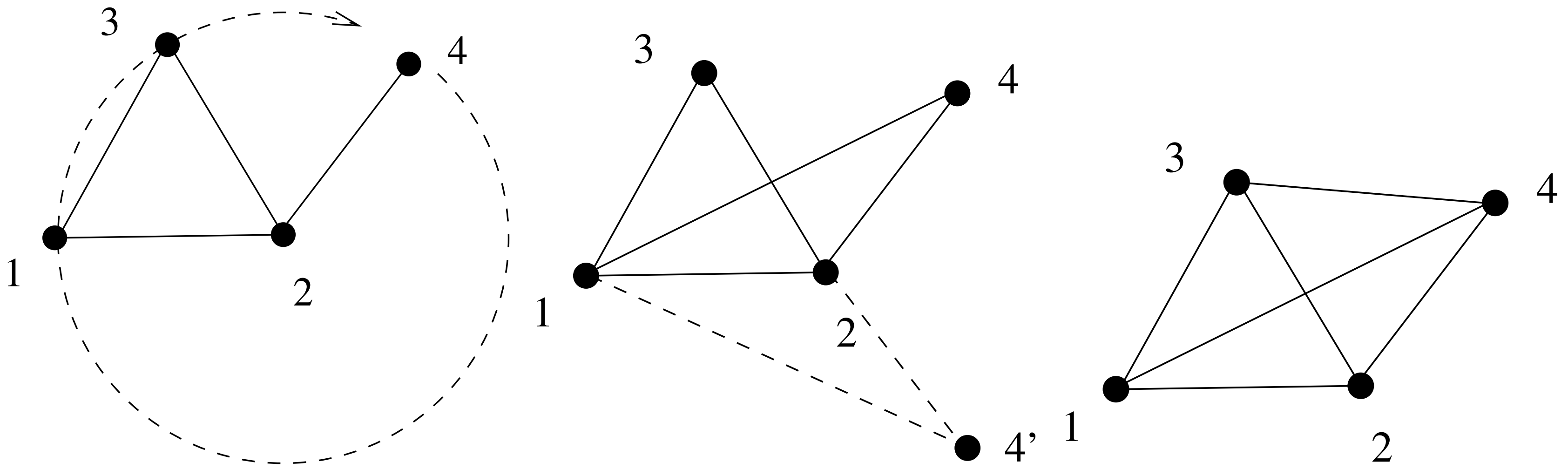
The number of embeddings

- Certain graphs have uncountably many (incongruent) embeddings
- Others have finitely many
- Cliques, for example, have at most one



The number of embeddings

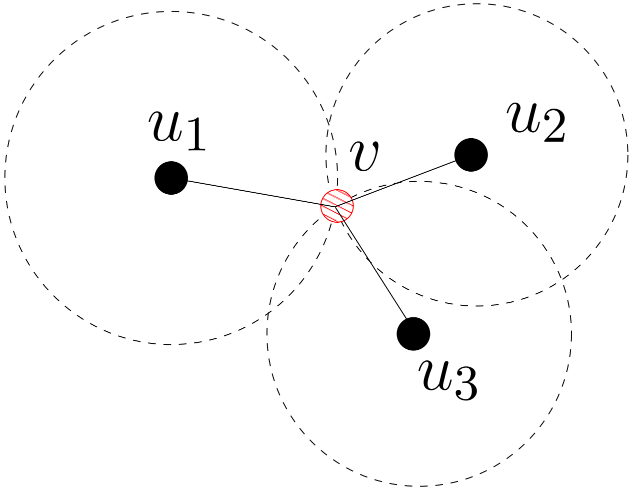
- Certain graphs have uncountably many (incongruent) embeddings
- Others have finitely many
- Cliques, for example, have at most one



Focus on discrete cases: get a combinatorial constraint problem with decision variables in continuous space

Vertex orders and embeddings

- Assume G has an embedding
- If $\exists$ an order $<$ on V such that:
  1. an embedding is known for the first $K+1$ vertices
  2. the v-th vertex is adjacent to at least $K+1$ predecessors
- Then x_v is the unique intersection of spheres $S(x_u, d_{uv})$ for u adjacent predecessor of v



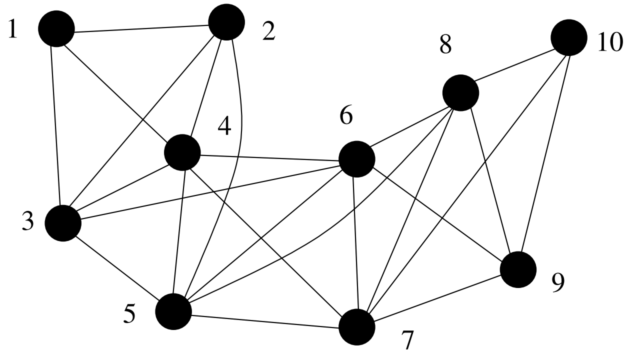
An interesting graph class

- So, if every $(K+1)$-tuple of consecutive vertices is a clique in G, we can find an embedding in polynomial time
- *(Computing a $K+1$ sphere intersection in $\mathbb{R}^K$ amounts to solving a square linear system)*

- Consider graphs with a weaker condition

every K-tuple of consecutive vertices is a clique in G

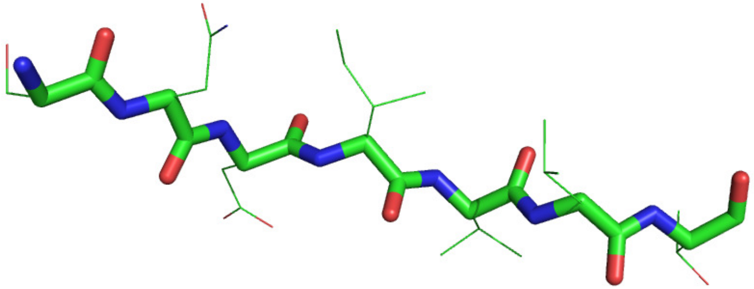
- This is called the DISCRETIZABLE MOLECULAR DISTANCE GEOMETRY PROBLEM (DMDGP)



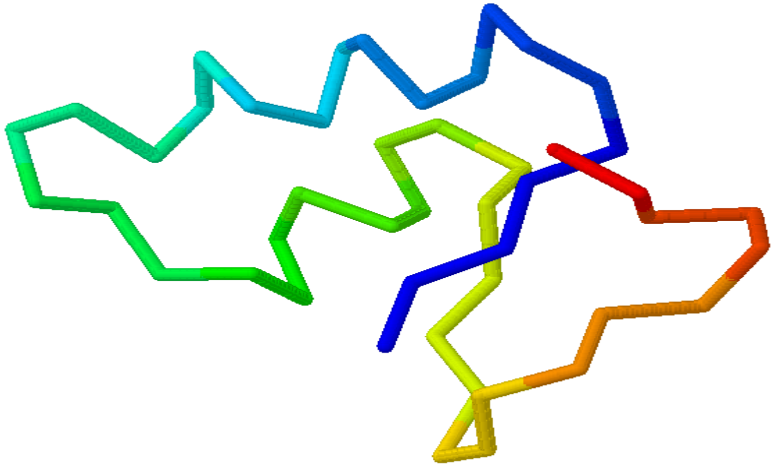
Instance with $K = 3$

Proteins

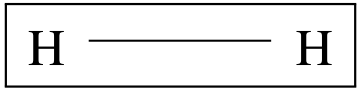
- Proteins are organized into a *backbone* and some *side chains*

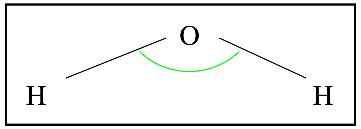


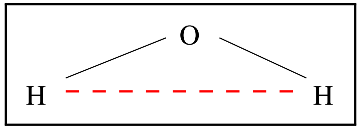
- Once the backbone is placed in $\mathbb{R}^3$, placing the side chains is known as the SIDE CHAIN PLACEMENT PROBLEM (SCPP) [Santana et al. ’08, Kim ’11]
- The backbone is a total order $<$ on a set V of atoms



Protein distances

- Covalent bond distances $d_{v-1,v}$ are known 

- Angles between covalent bonds are known 

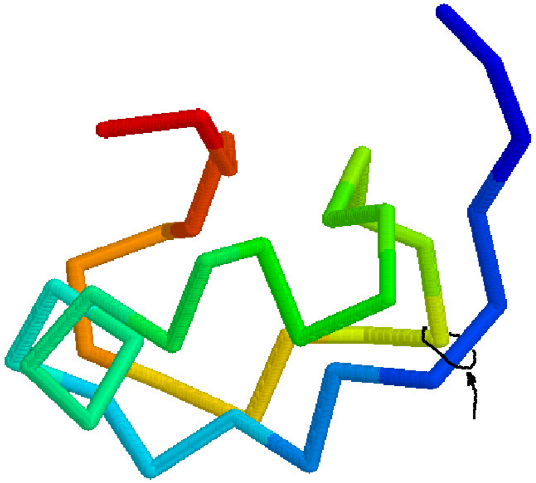
- $\Rightarrow d_{v-2,v}$ is known for all $v > 3$ 

- Distances $d_{v-3,v}$ are always < 6Å, so they can be measured using NMR techniques

We assume these distances are exact: this is false in practice, but we can find orders for which this assumption holds (see later if I have time)

- NMR might give other distances too

Atoms may be distant order-wise but closer than 6Å in space



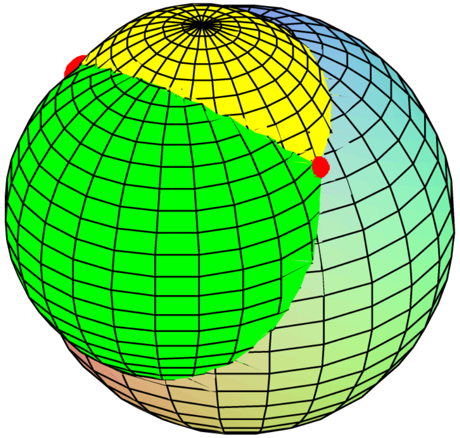
Sphere intersection

Situation:

- $x_{v-3}, x_{v-2}, x_{v-1}$ are known
- $d_{v,v-1}, d_{v,v-2}, d_{v,v-3}$ are known

and we're trying to find x_v

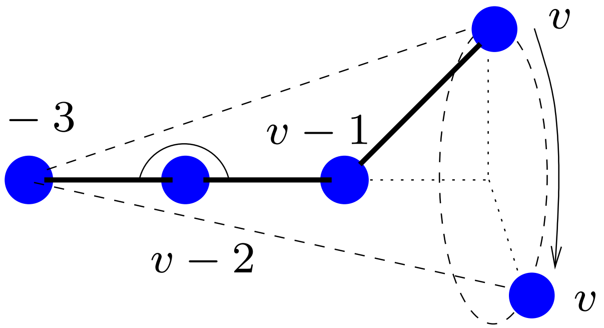
Then $x_v \in \bigcap_{i\in\{1,2,3\}} S(x_{v-i}, d_{v-i,v})$, the intersection of 3 spheres in $\mathbb{R}^3$, which in general contains 2 points



When does it fail?

The intersection of 3 spheres in $\mathbb{R}^3$ might fail to have *exactly two* points:

- it has zero points if the spheres do not intersect (but then the graph fails to have an embedding)
- it has uncountably many points (or a single one) if $d_{v-3,v-1} = d_{v-3,v-2} + d_{v-2,v-1}$



- Since the set of “flat triangles” over $v-3, v-2, v-1$ has Lebesgue measure 0 in the set of all triangles, this event has probability 0

The Branch-and-Prune algorithm

v: rank of current atom $x_{<v}$: partial embedding to rank $v-1$

G: instance X: current pool of embeddings

$S(y,r)$: $\mathbb{R}^K$ sphere centered at y with radius r

```
BranchAndPrune($v$, $x_{<v}$, $G$, $X$):
  Let $\mathcal{S} \leftarrow \bigcap_{i\in\{1,\ldots,K\}} S(x_{v-i}, d_{v-i,v}) = \{s_1,\ldots,s_q\}$, where $q \in \{0,2\}$
  for $s \in \mathcal{S}$ do
    Extend the current embedding to $x = (x_{<v}, s)$
    if $\forall u \in$ AdjPred$(v)$ $\|x_u - x_v\| = d_{uv}$ then
      if $(v = n)$ then
        Let $X \leftarrow X \cup \{x\}$
      else
        BranchAndPrune($v+1$, $x$, $G$, $X$)
      end if
    end if
  end for
```

BP properties

- The DMDGP is NP-hard [Lavor et al., COAP, to appear]
- The BP has worst-case exponential time
- With probability 1, it finds **all** incongruent embeddings of G extending the initial partial embedding known for $x_1, \ldots, x_K$
- In practice, it performs very efficiently with respect to speed and accuracy
- Can embed 10,000 vertices in a 13 seconds of CPU time
- Two empirical observations:
  1. **the number of solutions it finds is always a power of two**
  2. $|V|$ **versus CPU time plots are always linear-like for PDB**

Symmetry

BP root node symmetry

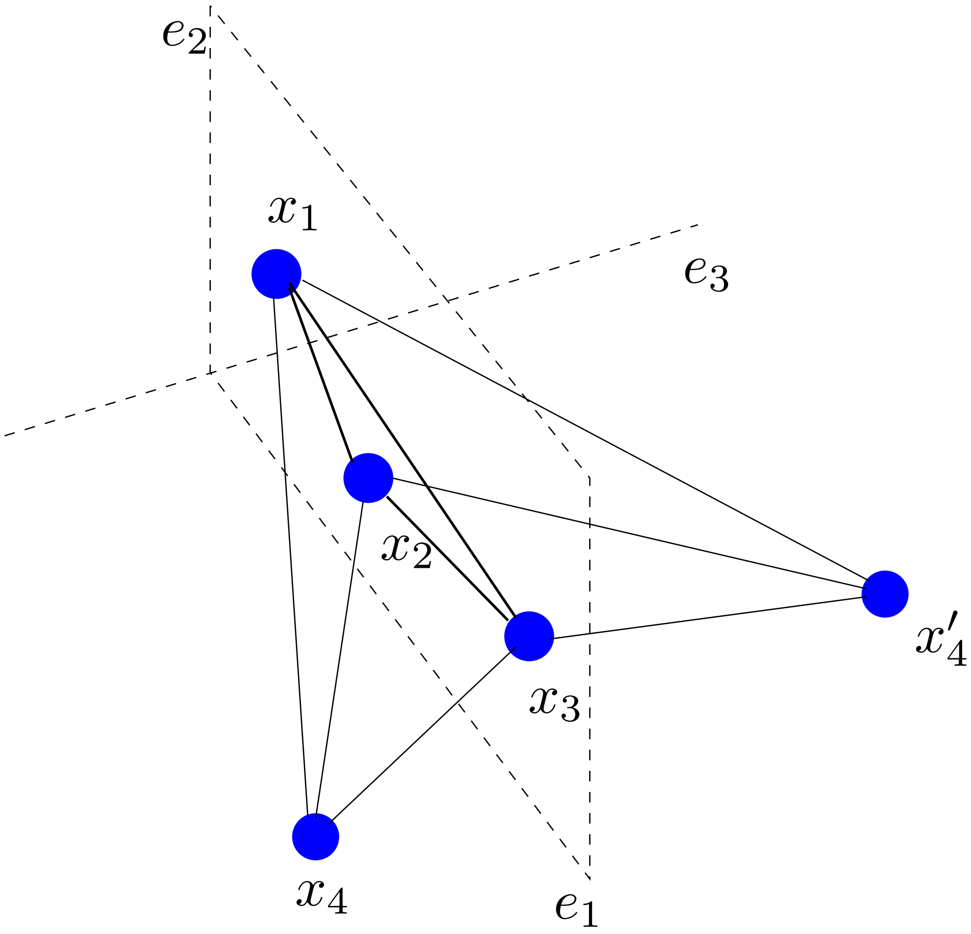
[Lavor et al. COAP, to appear]

- Once the first 3 atoms are placed, the fourth can generally be placed in two positions x_4, x'_4

Thm.

- x'_4 is a reflection of x_4 w.r.t. the plane defined by x_1, x_2, x_3

- The BP tree is symmetric below level 3, so it suffices to just consider half of the BP tree



Number of solutions

Instance	$\lvert X\rvert$
`mmorewu-2`	2
`mmorewu-3`	2
`mmorewu-4`	4
`mmorewu-5`	4
`mmorewu-6`	4
`lavor10_0`	4
`lavor15_0`	16
`lavor20_0`	8
`lavor25_0`	8
`lavor30_0`	2
`lavor35_0`	64
`lavor40_0`	2
`lavor45_0`	2
`lavor50_0`	4096
`lavor55_0`	64
`lavor60_0`	64

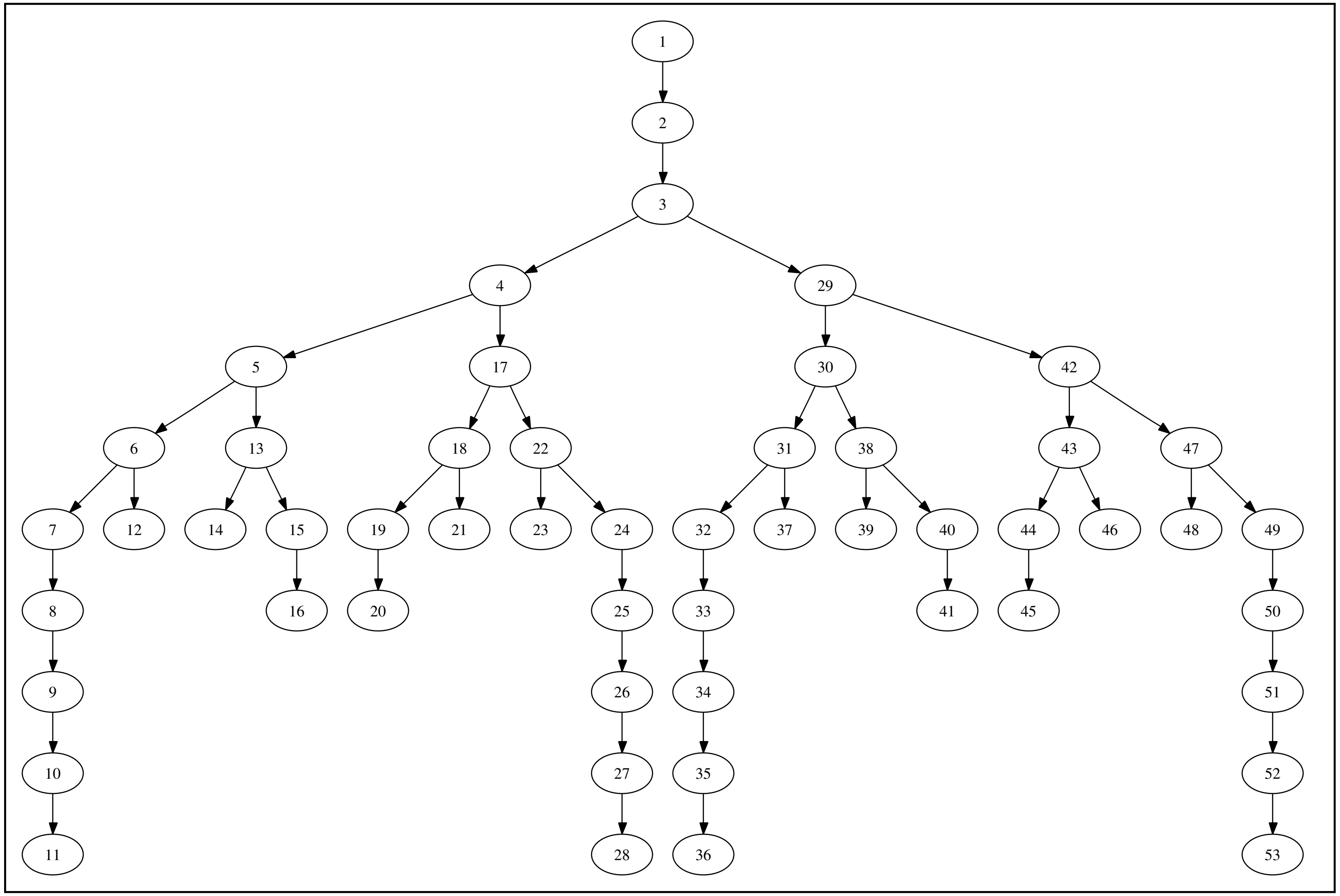
Instance	$\lvert X\rvert$
`1brv`	1
`1aqr`	2
`2erl`	1
`1crn`	1
`1ahl`	8
`1ptq`	1
`1brz`	2
`1hoe`	1
`1lfb`	1
`1pht`	1
`1jk2`	1
`1f39a`	1
`1acz`	4
`1poa`	1
`1fs3`	1
`1mbn`	1
`1rgs`	1
`1m40`	1
`1bpm`	1
`1n4w`	1
`1mqq`	1
`1rwh`	1
`3b34`	1
`2e7z`	1
`1epw`	1

For all tested DMDGP instances, $\exists \ell \in \mathbb{N}$ such that $|X| = 2^\ell$

← results only refer to $\frac{1}{2}$ of the tree, multiply by 2 to get $|X|$

A BP search tree example

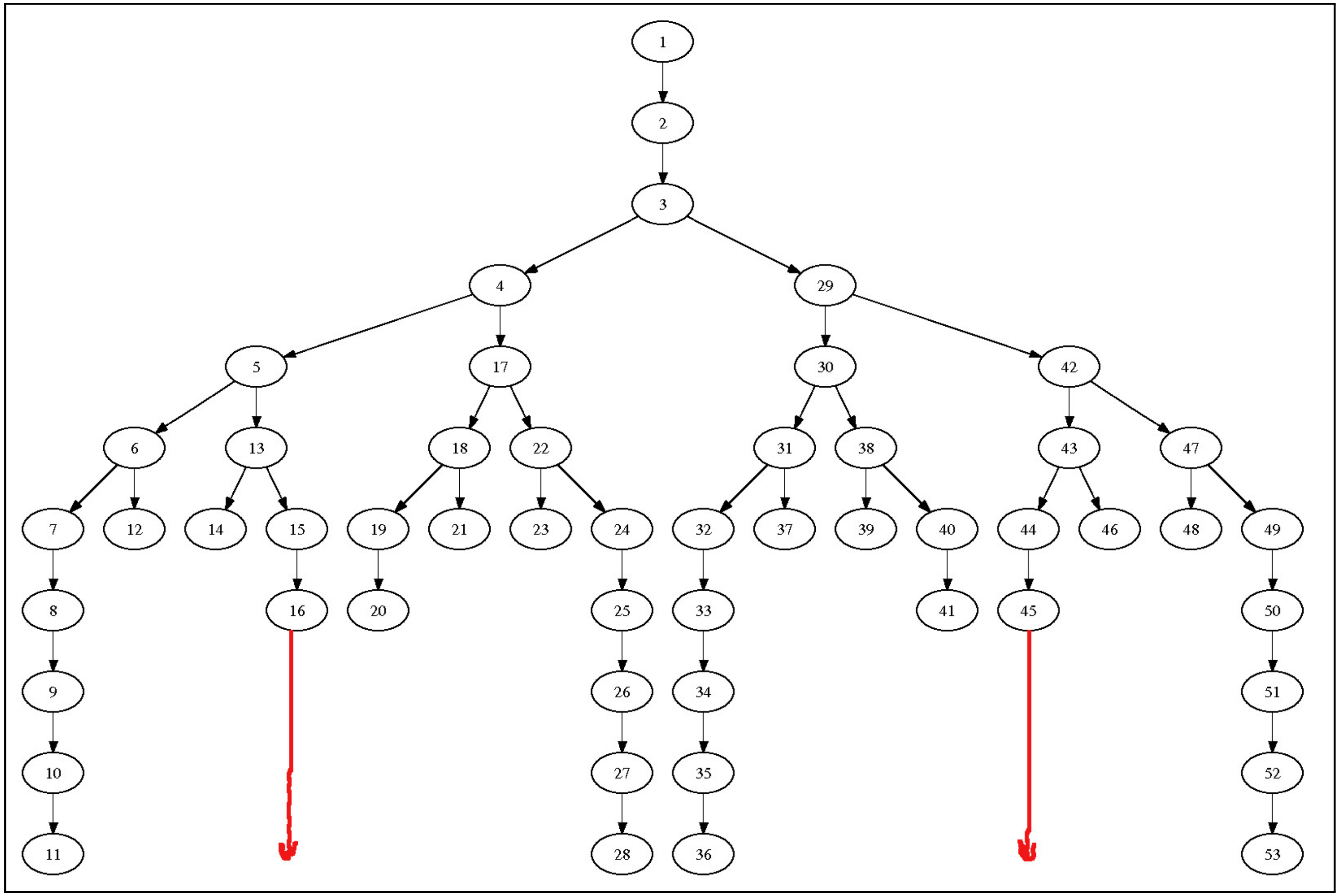
- Typical BP search tree (embeddings = paths root→leaves)



- Root node symmetry forces $|X|$ to be even
- No evident reason why $|X|$ should be a power of two

A BP search tree example

- Typical BP search tree (embeddings = paths root→leaves)



- Root node symmetry forces $|X|$ to be even
- No evident reason why $|X|$ should be a power of two (why not symmetric paths to level $|V|$ from nodes 16 and 45?)

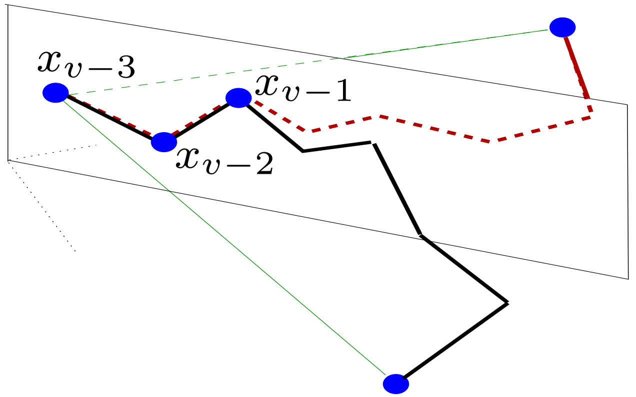
Discretization/pruning distances

- Let $E_D = \{\{u,v\} \mid |u-v| \le K\}$ and $E_P = E \smallsetminus E_D$
- E_D are the **discretization distances**
  - they guarantee that the instance is a DMDGP
  - they allow the construction of the complete BP tree
  - this tree has $2^{|V|-3}$ leaves, $2^{|V|-4}$ if we consider root node symmetry
- E_P are the **pruning distances**
  - they allow pruning of the BP tree
  - not at all clear why they should prune branches symmetrically

Symmetry by pruning distances

[Liberti et al., LNCS (COCOA), 2011]

Given an embedding x, let R_x^v be the reflection w.r.t. the hyperplane through $x_{v-K}, \ldots, x_{v-1}$



Thm.

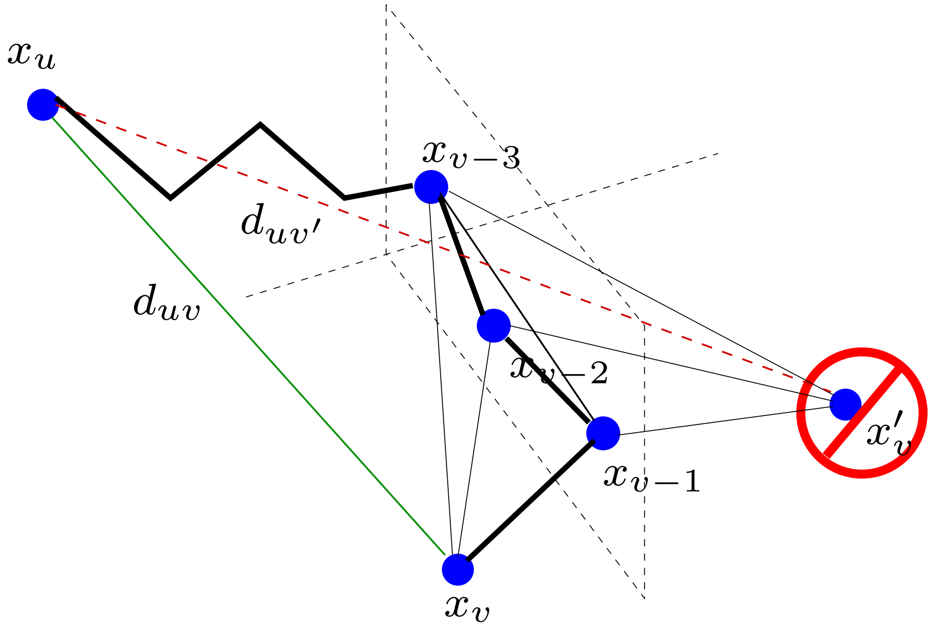
With prob. 1, for all $v > K, u < v - K$ there is a finite set $H^{uv} \subseteq \mathbb{R}_+$ with $|H^{uv}| = 2^{v-u-K}$ s.t.

$$\forall x \in X \; (\qquad \underbrace{\|x_u - x_v\|}_{\text{plays the role of pruning dist.}} \qquad \in H^{uv})$$

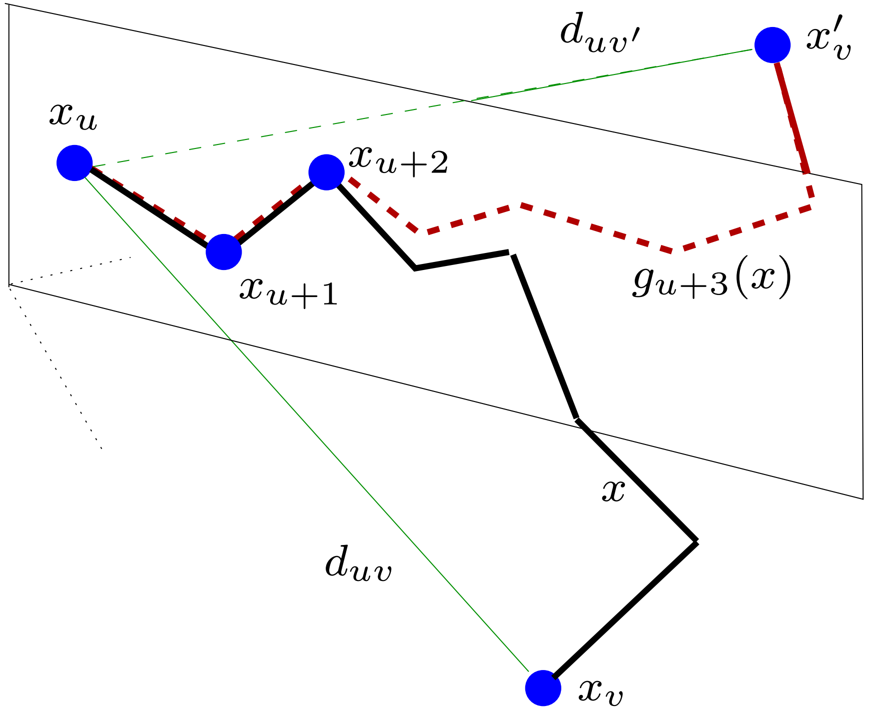
Furthermore, for $x' \in X \smallsetminus \{x\}$,

$$\|x_u - x_v\| = \|x'_u - x'_v\| \text{ iff } x'_v = R_x^{u+K}(x_v)$$

Reflection symmetry



“Root symmetry” may fail when not at root

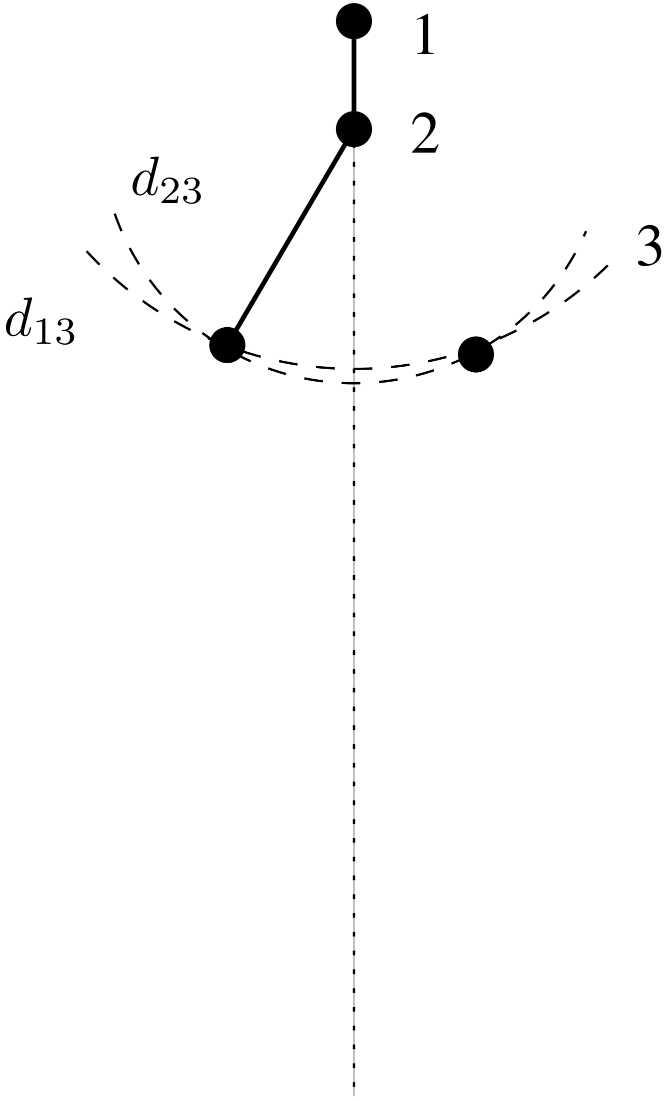


But then ∃ another symmetry

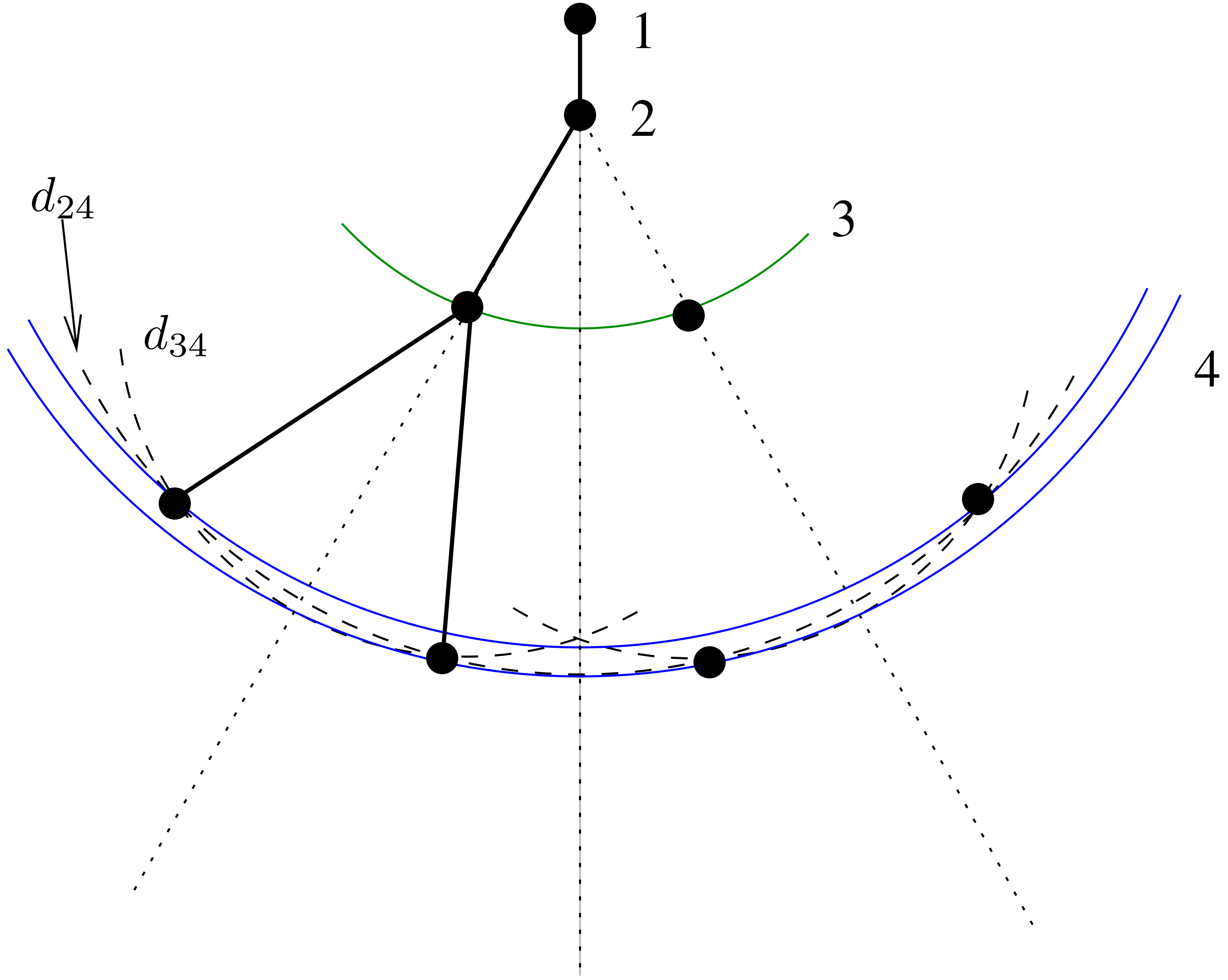
Reflection “from rank v”: define *partial reflections* operators

$$g_v(x) = (x_1, \ldots, x_{v-1}, R^v_x(x_v), \ldots, R^v_x(x_n)) \quad (2)$$

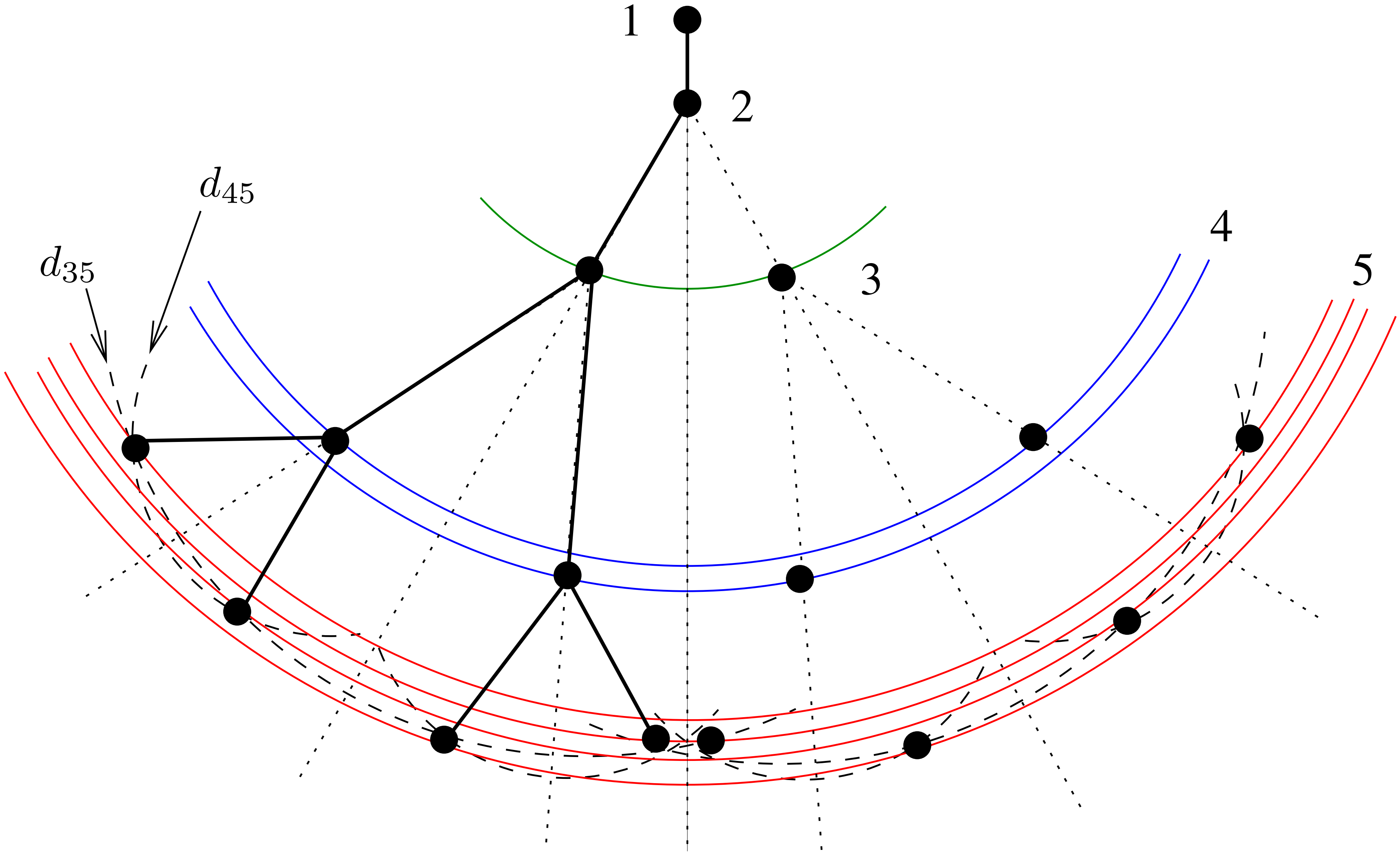
Structure of the BP tree ($\mathbb{R}^2$)



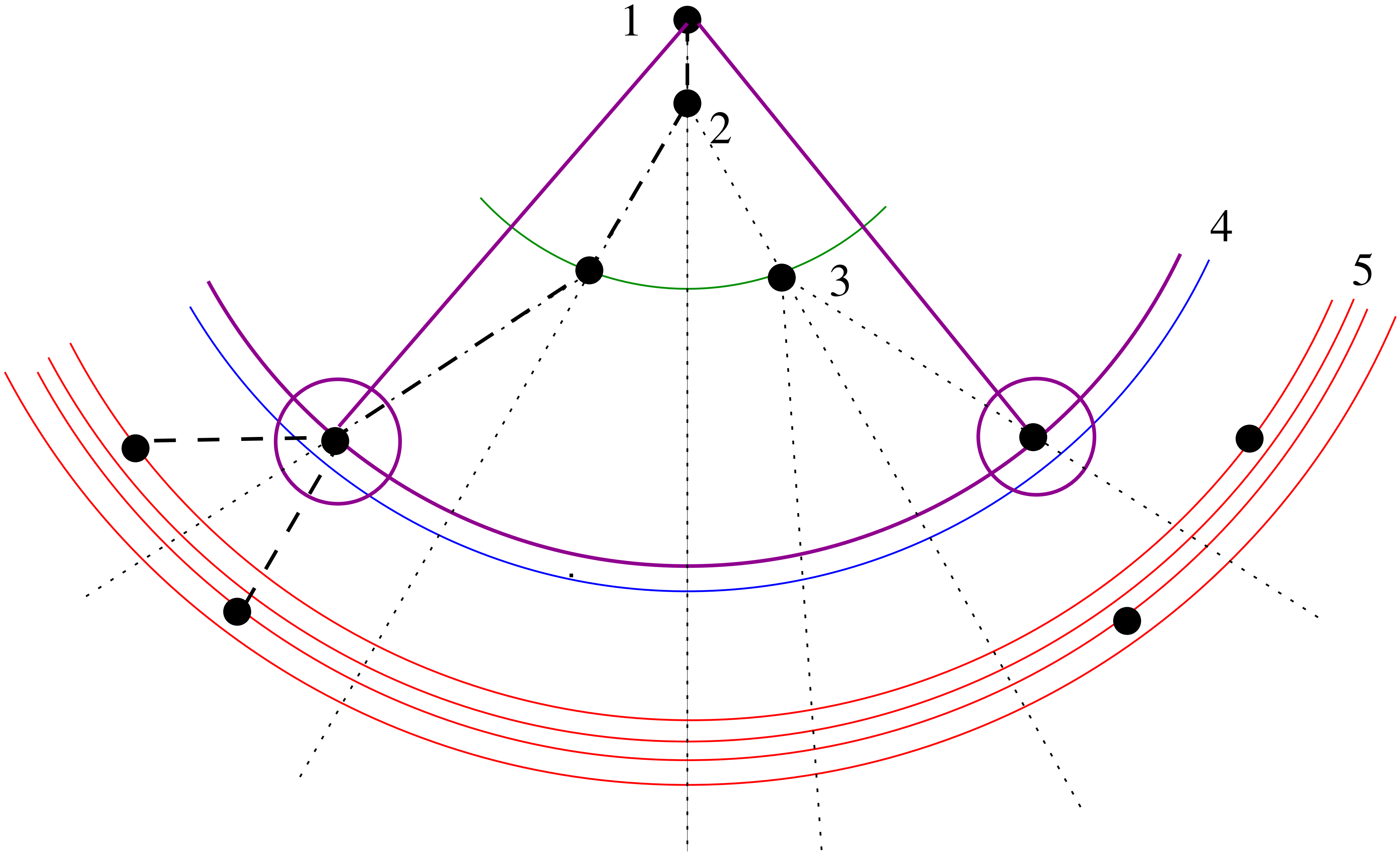
Structure of the BP tree ($\mathbb{R}^2$)



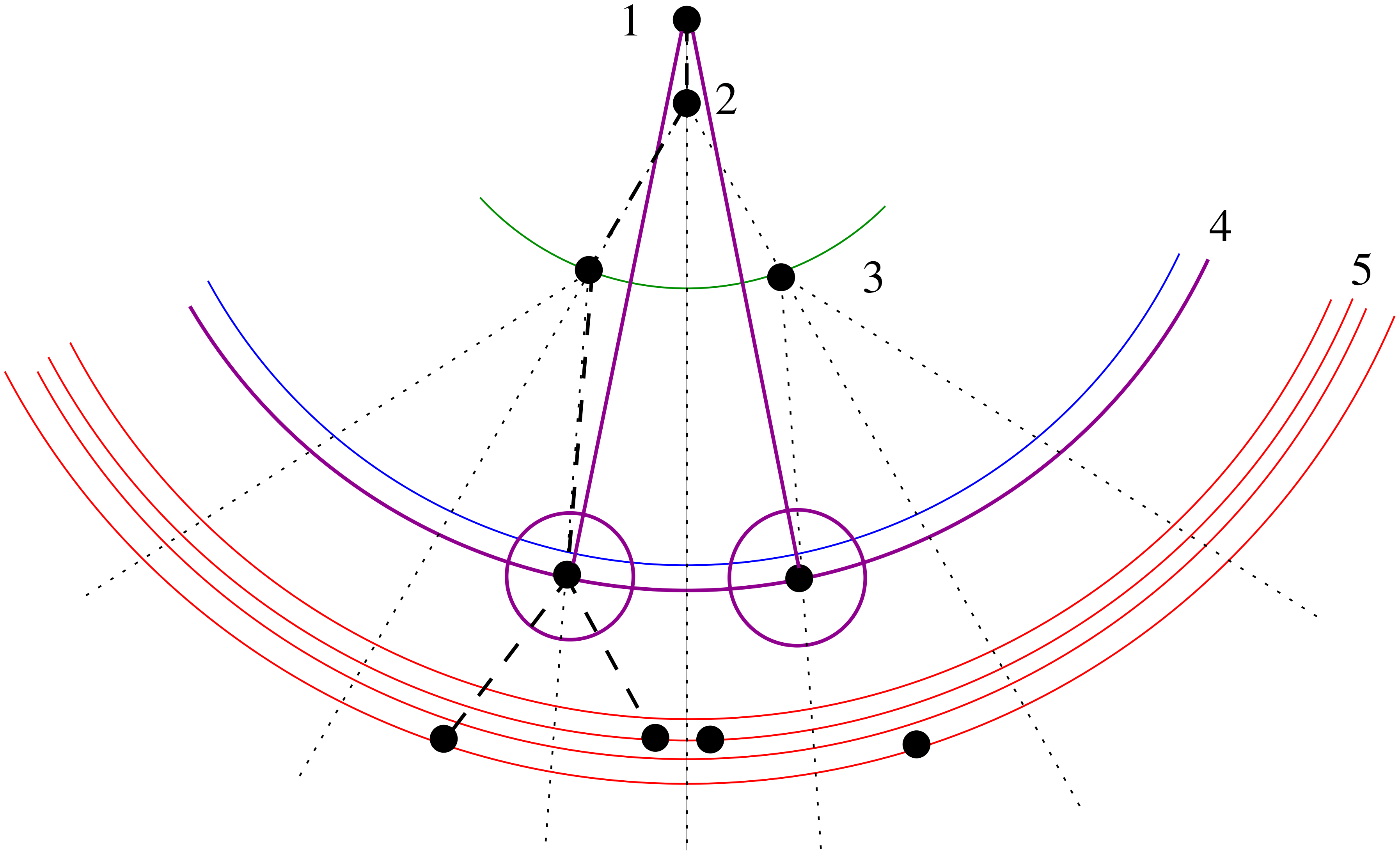
Structure of the BP tree ($\mathbb{R}^2$)



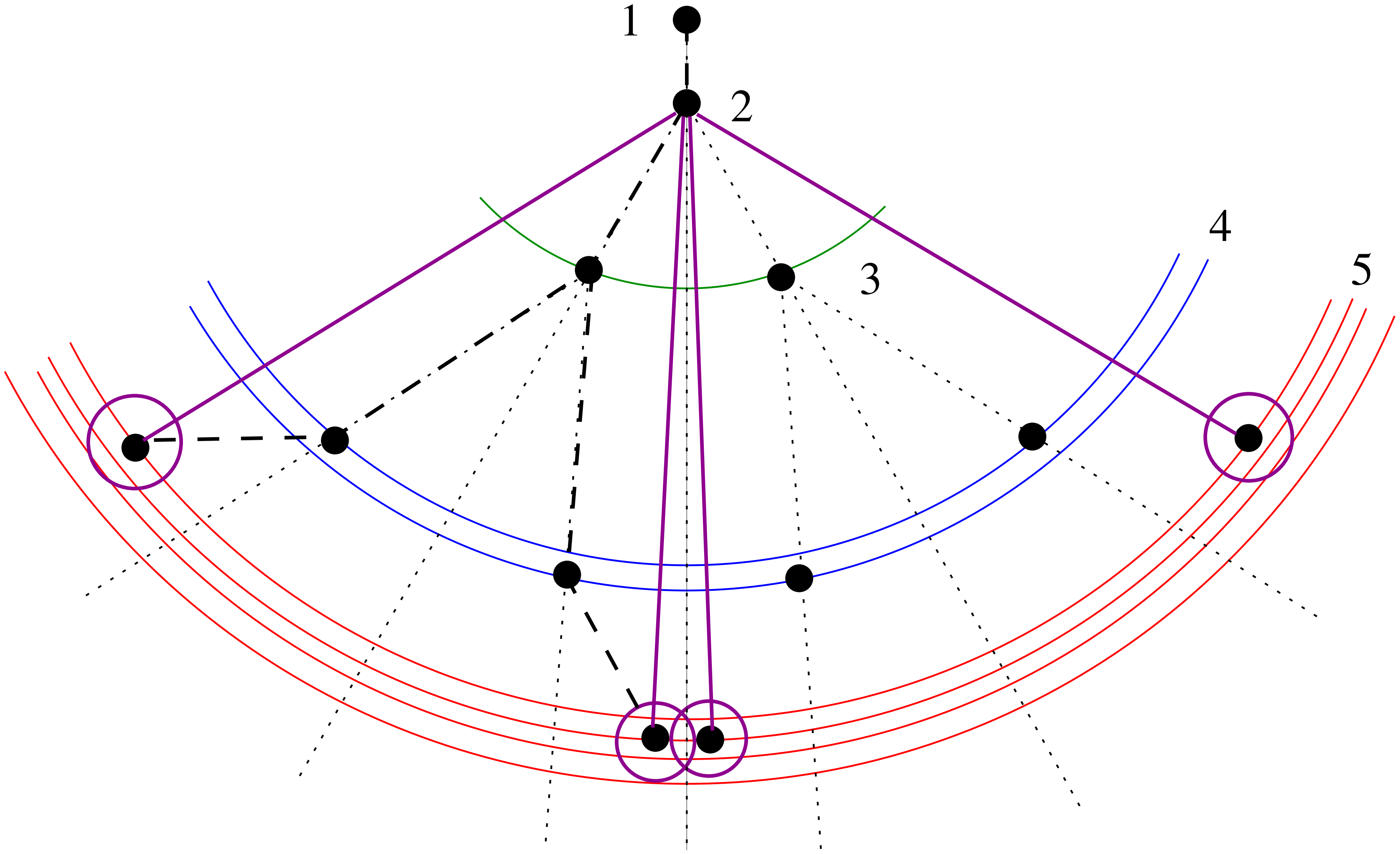
Effect of pruning distance d_{14}



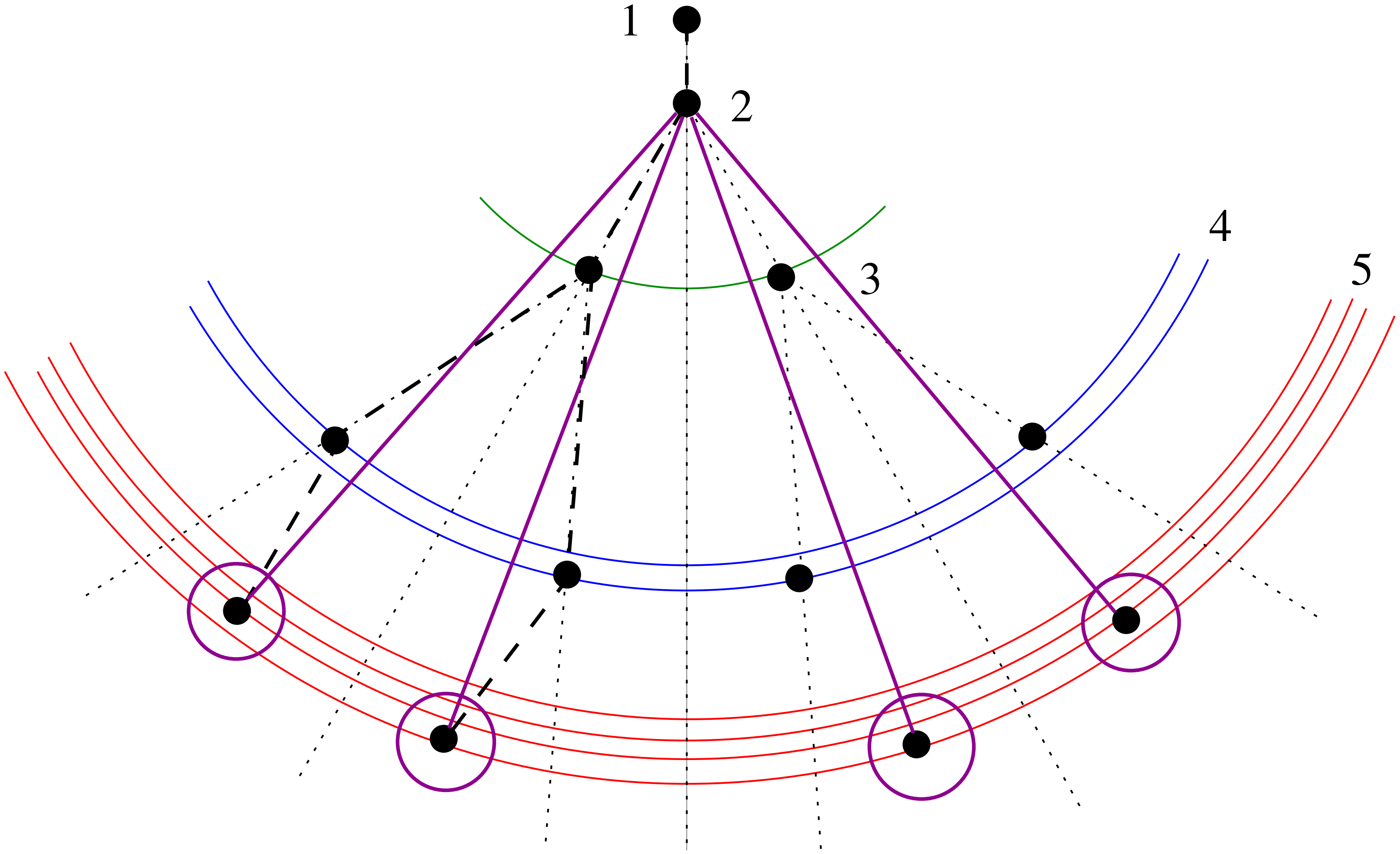
Effect of pruning distance d_{14}



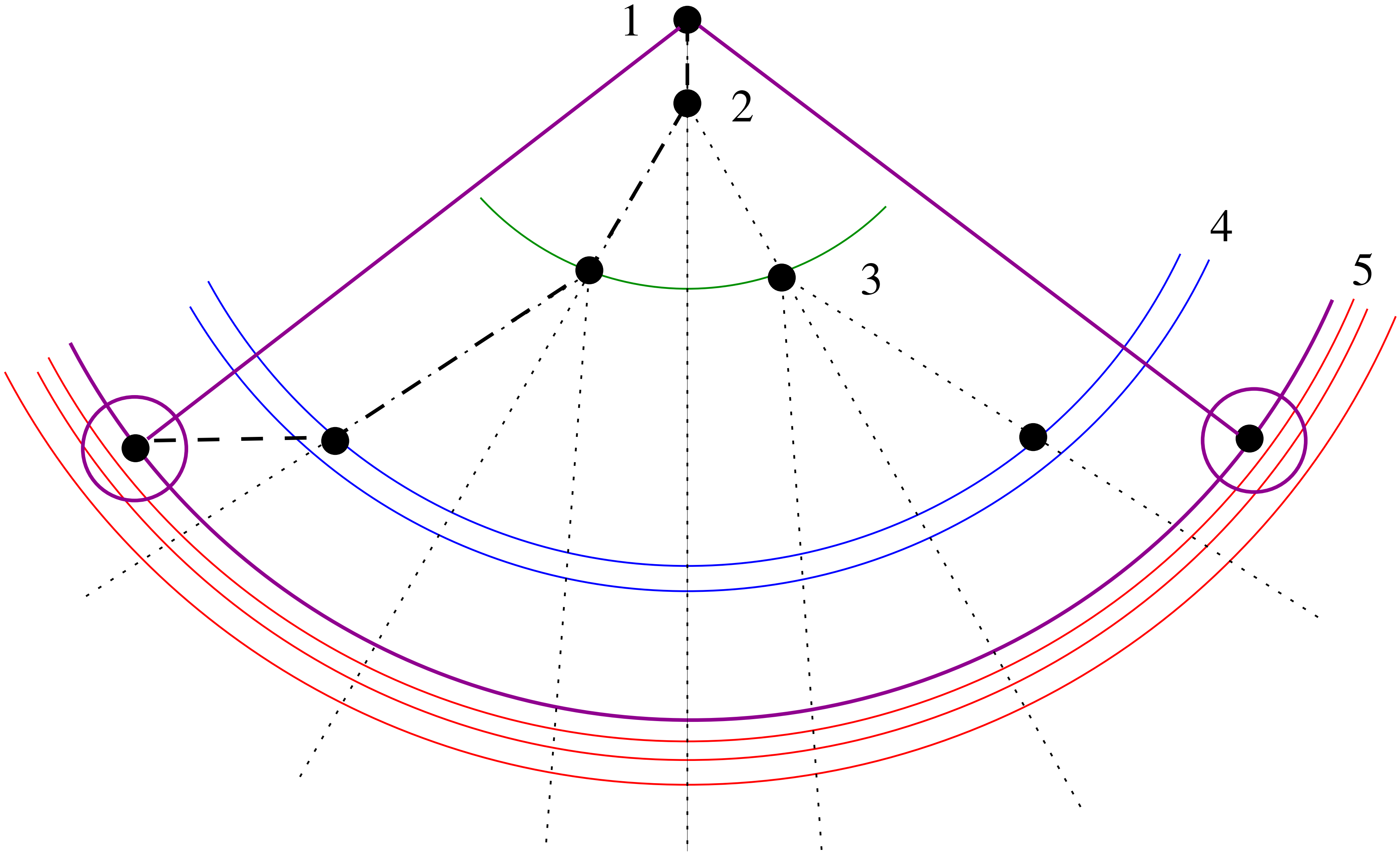
Effect of pruning distance d_{25}



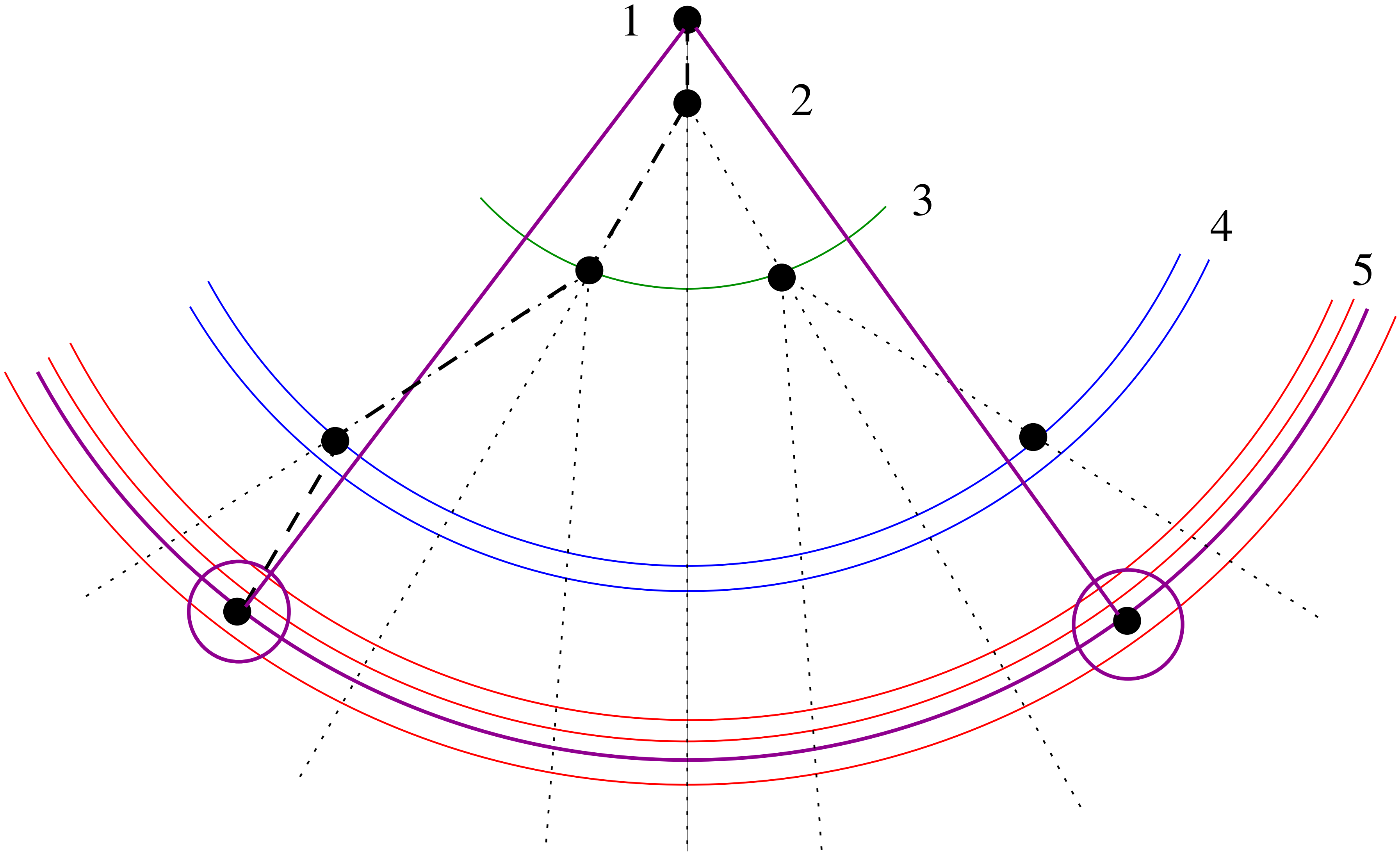
Effect of pruning distance d_{25}



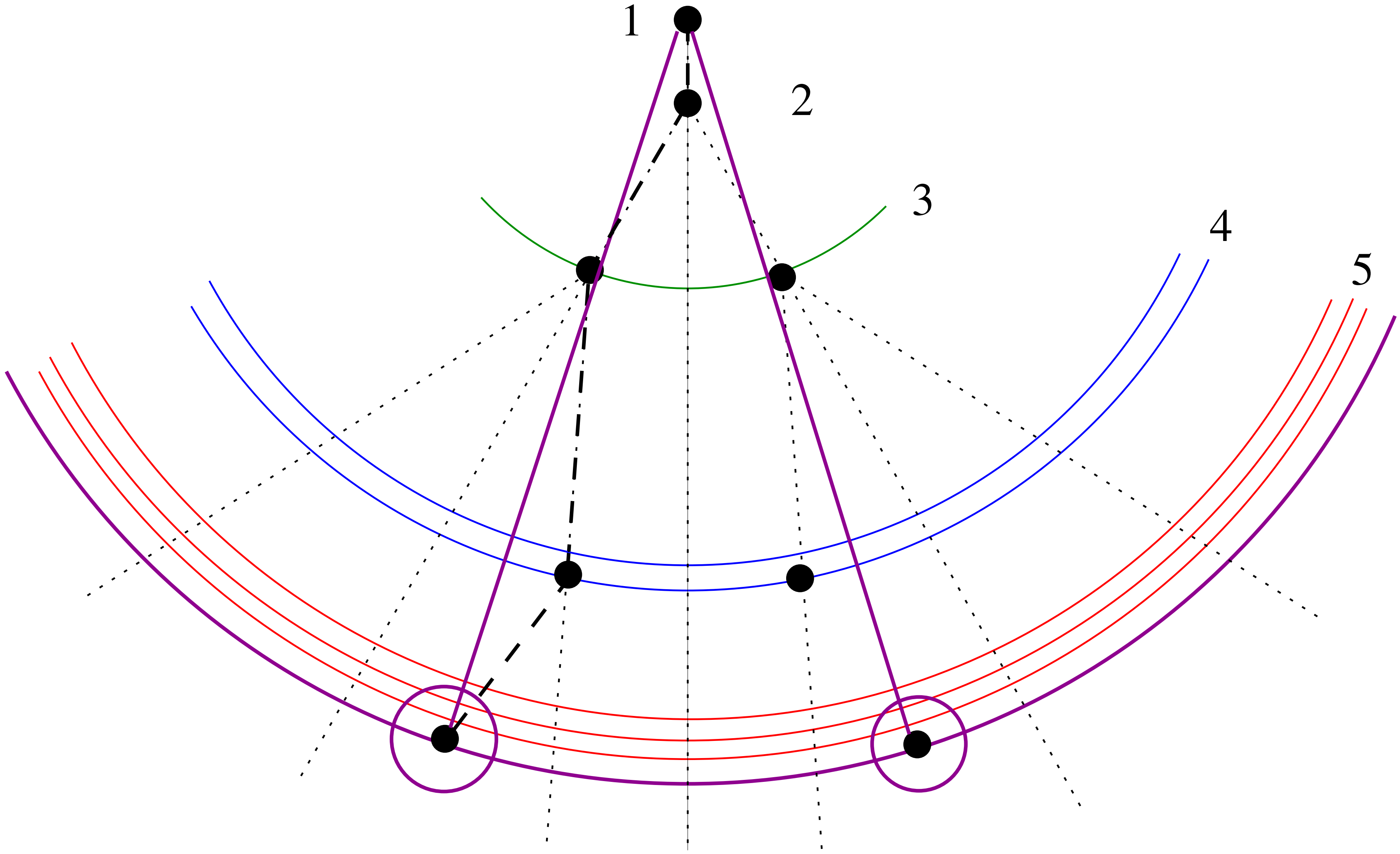
Effect of pruning distance d_{15}



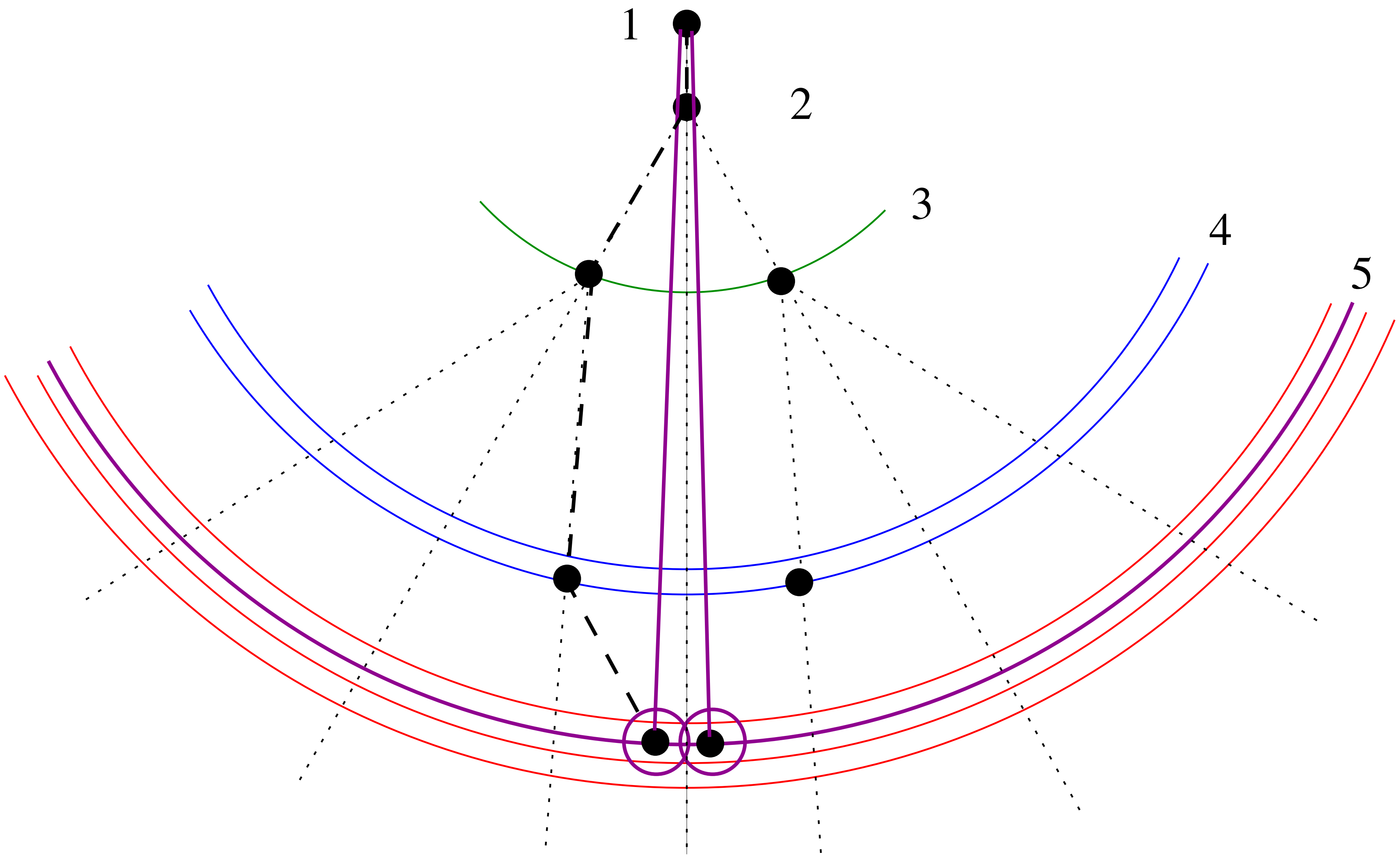
Effect of pruning distance d_{15}



Effect of pruning distance d_{15}

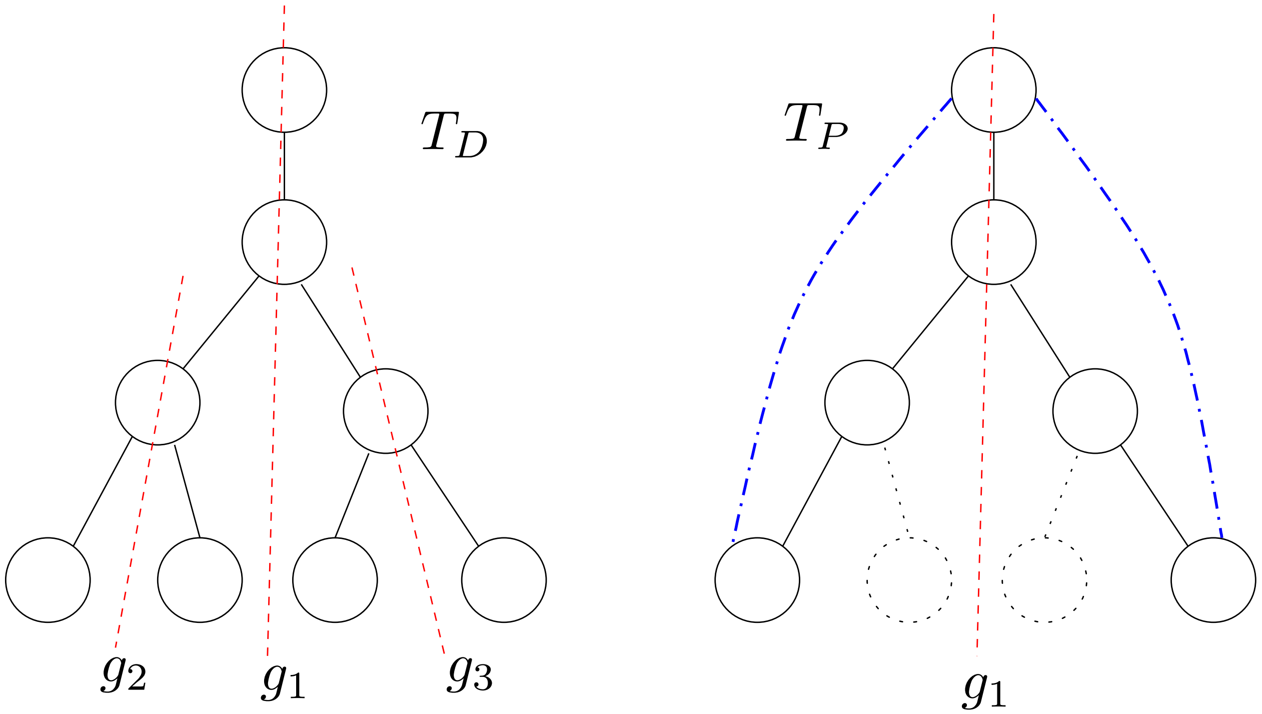


Effect of pruning distance d_{15}



Groups fixing the trees

- Let T_D be a full BP binary search tree
- Let T_P be the subtree of T_D representing only feasible branches
- Draw them so $T_P \subseteq T_P$
- Invariant group for T_D: all partial reflections (g_1, g_2, g_3)
- Invariant group for T_P: only some partial reflections (g_1)



Discretization group

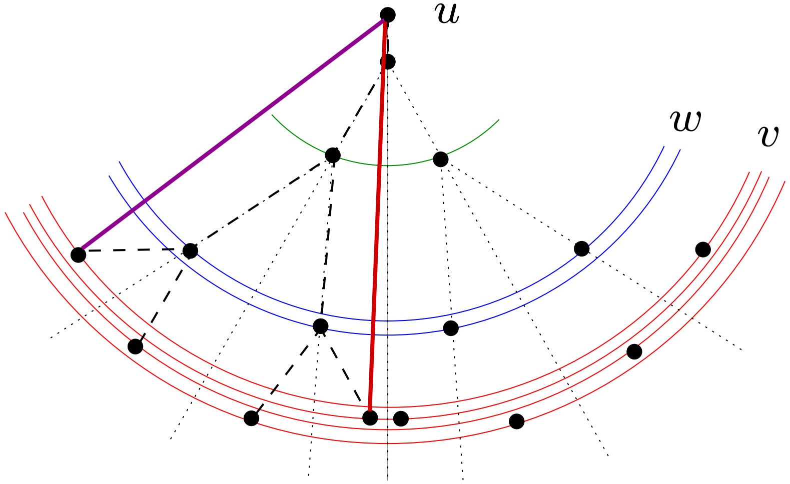
Group of partial reflections fixing the complete BP tree (no pruning distances)

- The following hold with probability 1 $\forall v > K$:
  1. g_v is injective with probability 1 (by reflection)
  2. $\boxed{g_v \text{ is idempotent}}$ (by reflection)
  3. $\forall u > K, u \neq v$, g_u and g_v commute (nontrivial)
- Thus, $\mathcal{G}_D = \langle g_v \mid v > K \rangle$ is an Abelian group under composition $\Rightarrow$ $\boxed{\text{isomorphic to } C_2^{n-K})}$
- By previous thm, discretization distances are invariant under $\mathcal{G}_D$
- The action of $\mathcal{G}_D$ on X is transitive, i.e. $\forall x, x' \in X \exists g \in \mathcal{G}_D \ (x' = g(x))$
- This action has only one orbit, i.e. $X = \mathcal{G}_D x$

Pruning group

Group of partial reflections fixing the actual BP tree (with pruning distances)

- Assume DMDGP instance is YES, consider $\{u,v\} \in E_P$
- With probability 1, $d_{uv} \in H^{uv}$ (otherwise the instance would be NO)
- Notice $d_{uv} = \|x_v - x_u\| \neq \|g_w(x)_v - g_w(x)_u\|$ for all $w \in \{u+K+1,\ldots,v\}$



- In order to keep invariance we remove such g_w's from the group
- **Pruning group**: $\mathcal{G}_P = \langle g_w \mid w > K \wedge \forall \{u,v\} \in E_P \ (w \notin \{u+K+1,\ldots,v\}) \rangle$
- $\boxed{\mathcal{G}_P \leq \mathcal{G}_D}$ and all distances are invariant w.r.t. the pruning group
- Again, $\boxed{\text{action of } \mathcal{G}_P \text{ on } X \text{ is transitive}}$ (nontrivial proof)

Power of two

Thm.

$\exists \ell \in \mathbb{N}\ (|X| = 2^{\ell})$

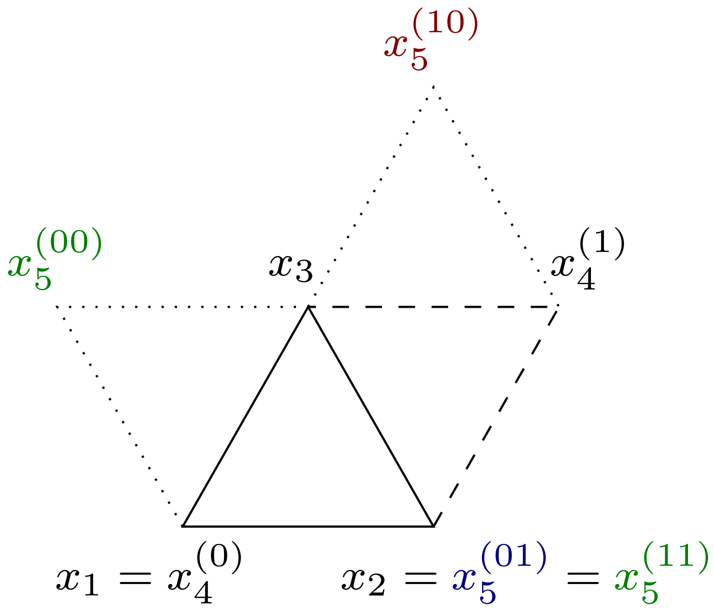
Proof

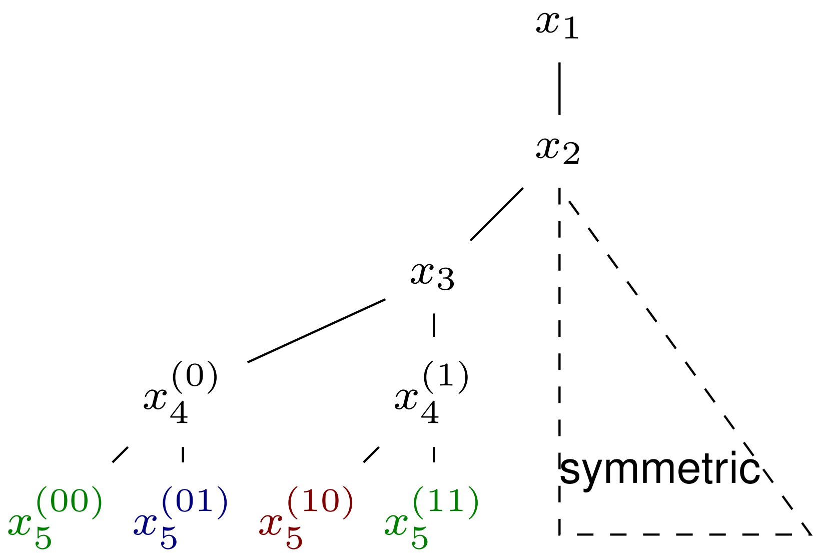
With probability 1:

- $\mathcal{G}_D \cong C_2^{n-K} \Rightarrow |\mathcal{G}_D| = 2^{n-K}$
- $\mathcal{G}_P \leq \mathcal{G}_D \Rightarrow |\mathcal{G}_P| \mid |\mathcal{G}_D| \Rightarrow \exists \ell \in \mathbb{N}\ |\mathcal{G}_P| = 2^{\ell}$
- Action of $\mathcal{G}_P$ on X is transitive $\Rightarrow \mathcal{G}_P x = X$
- Idempotency $\Rightarrow$ for $g, g' \in \mathcal{G}_P$, if $gx = g'x$ then $g = g' \Rightarrow |\mathcal{G}_P x| = |\mathcal{G}_P|$
- Thus, $|X| = |\mathcal{G}_P x| = |\mathcal{G}_P| = 2^{\ell}$

Why the “probability 1”?

- Not all “YES” DMDGP instances have $|X| = 2^\ell$
- But the set of such instances (with real data) has Lebesgue measure zero in the set of all DMDGP instances





Happens when > 1 vertices are embedded in the same position

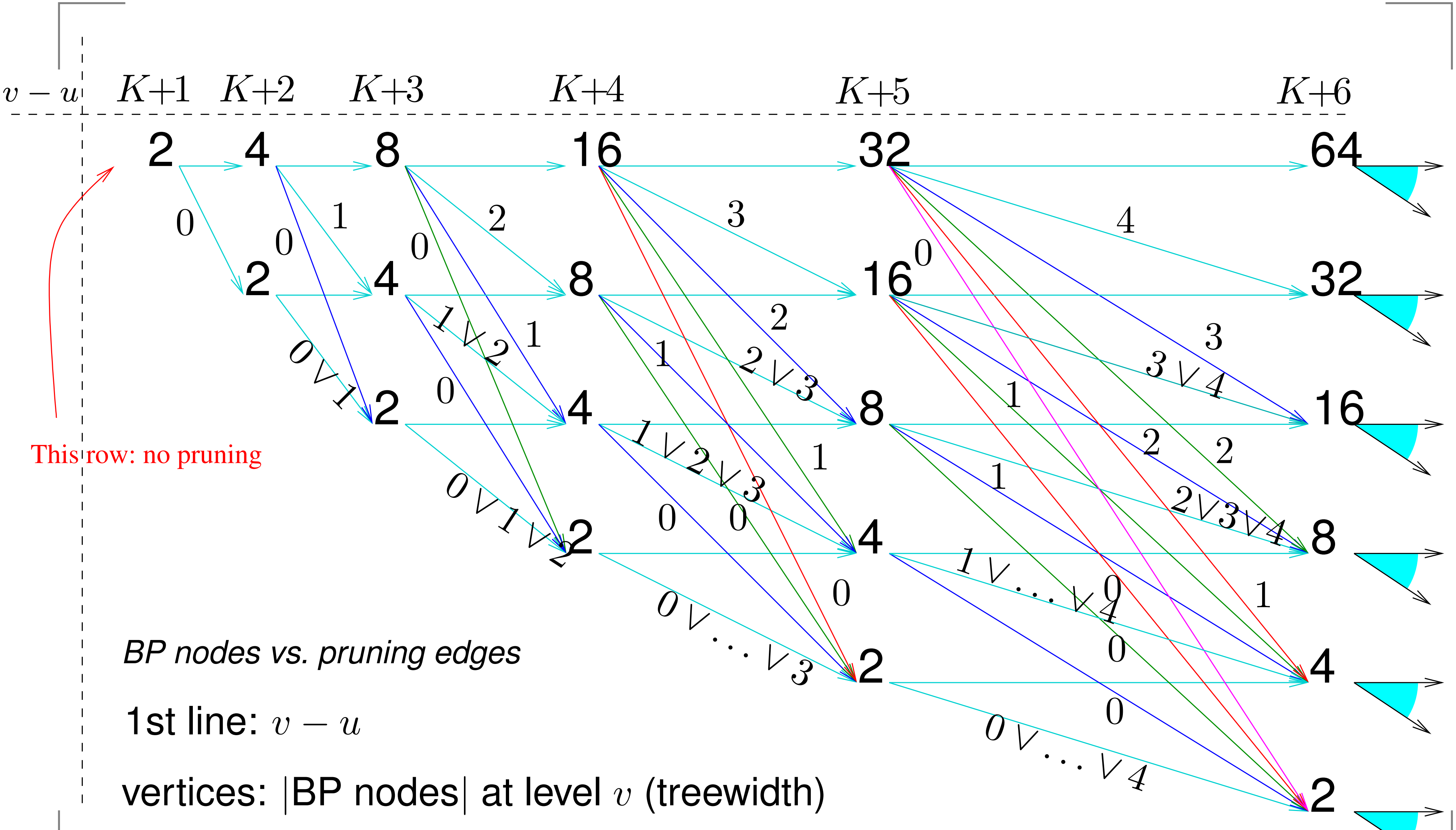
$x_5^{(01)}$ **should** be infeasible, but $x_5^{(01)} = x_5^{(11)}$ (event with prob. 0)

Polynomial cases

A polynomial BP?

- We never noticed any exponential-time increase behaviour in all our experiments (several scores of instances generated from PDB files)
- We recently embedded a 10000-atom protein backbone in 13s on one core
- It is easy to show that BP has worst-case exponential complexity
- Are a polynomial case of the DMDGP?
- Complexity depends on BP nodes; since height$\le |V|$, only need to consider treewidth
- A pruning edge $\{u, v\}$ with $u < v - K$ reduces the number of nodes at level v from 2^{v-K} to $2^{v-K-u+1}$ (by symmetry)

BP subtree rooted at u



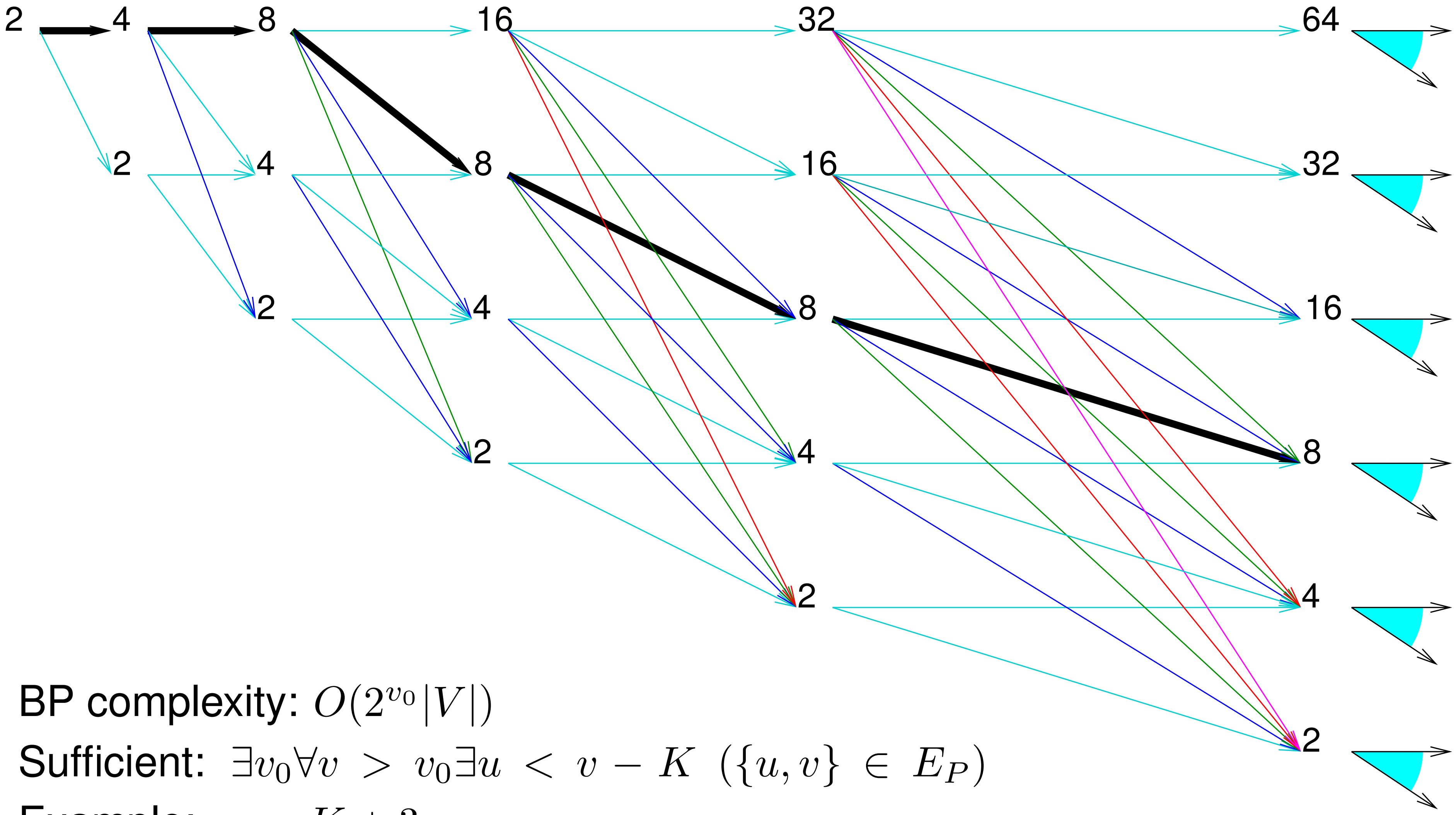
BP nodes vs. pruning edges

1st line: $v-u$

vertices: |BP nodes| at level v (treewidth)

arcs: $\exists$ pruning edge $\{u + \text{arc\_label}, v\}$

Constant treewidth

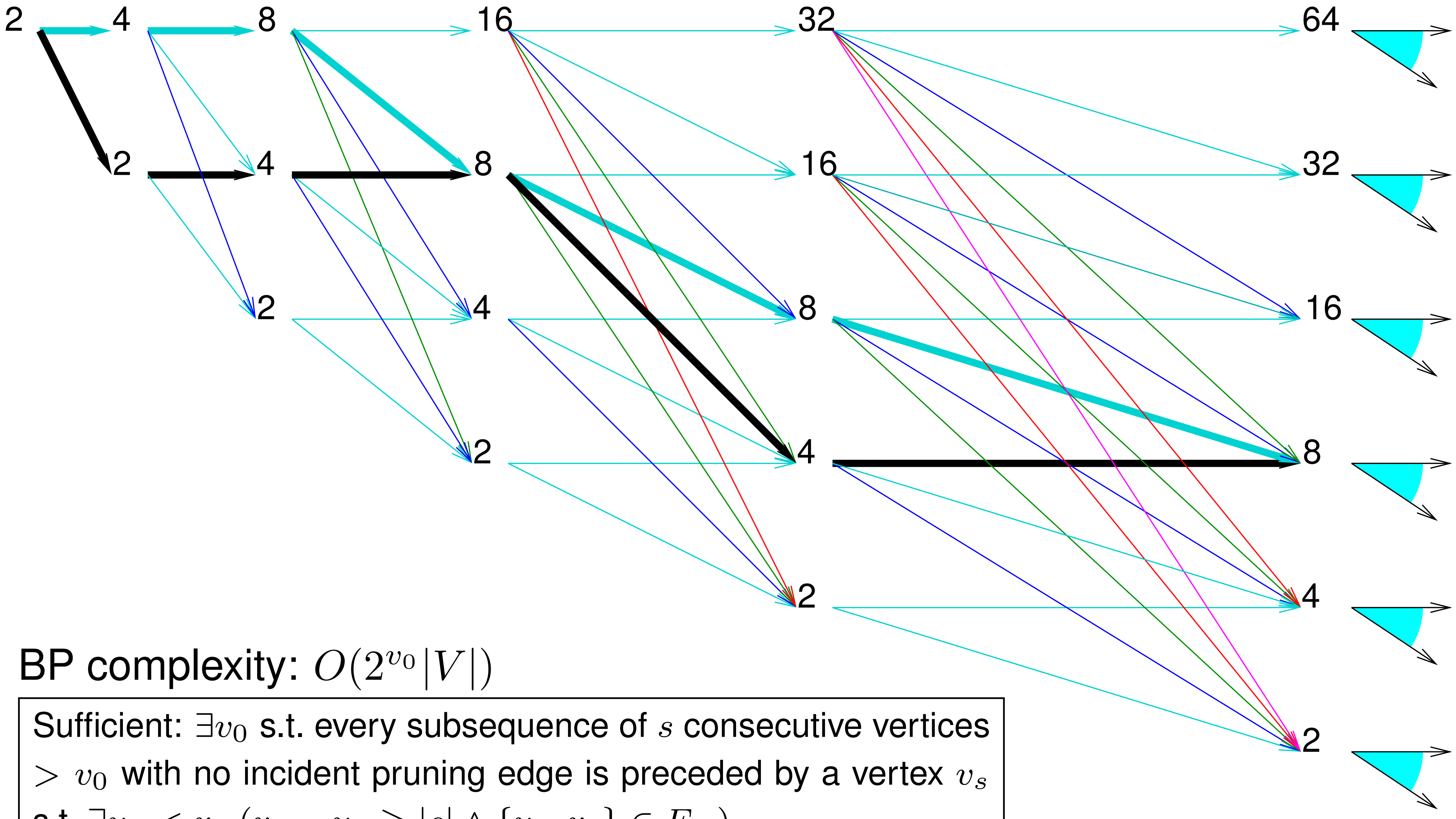


BP complexity: $O(2^{v_0}|V|)$

Sufficient: $\exists v_0 \forall v > v_0 \exists u < v - K \ (\{u, v\} \in E_P)$

Example: $v_0 = K + 3$

Constant-bounded treewidth



BP complexity: $O(2^{v_0}|V|)$

Sufficient: $\exists v_0$ s.t. every subsequence of s consecutive vertices $> v_0$ with no incident pruning edge is preceded by a vertex v_s s.t. $\exists u_s < v_s \ (v_s - u_s \geq |s| \wedge \{u_s, v_s\} \in E_P)$

“Any path under the constant path”

Polynomial time BP

- We can also allow treewidth growth as long as it's logarithmic in n
- This yields a polynomial-time BP

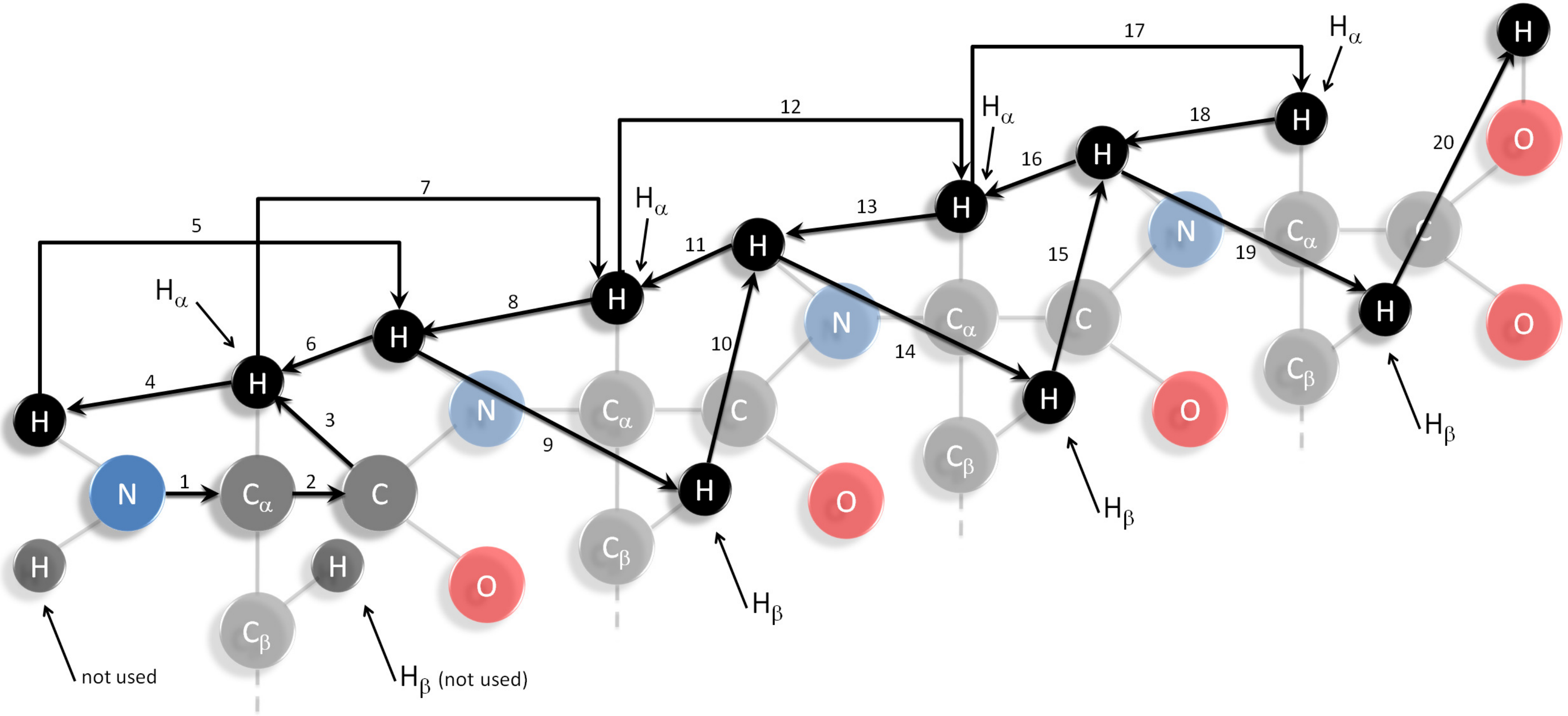
  . . . well, fixed-parameter tractable w.r.t. v_0

We tested all our protein instances: all display either constant or const-bounded treewidths **with very low** v_0 (i.e. $v_0 = 4$)

Application to proteomics

Virtual hydrogen backbone

- The most accurate NMR distances are **between hydrogen atoms only**, but the actual backbone is a chain of N-C_α-C groups
- So find a *virtual* backbone composed of hydrogens only, and such that its order satisfies the DMDGP requirements



Certain hydrogens must be enumerated twice

[Lavor et al. JOGO]

Listing atoms twice

- If a hydrogen is listed twice, then there are $i \neq j \in V$ indexing the same atom
- Thus $x_i = x_j$ and $d_{ij} = 0$
- For all k such that $\{i, k\} \in E$, we have that $\{j, k\} \in E$ as $d_{jk} = d_{ik} + 0$, and

$$d_{ij} + d_{jk} = 0 + d_{jk} = d_{ik}$$

so STRICT TRIANGULAR INEQUALITIES do not hold for all atom triplets

- However, it only fails on *nonconsecutive* triplets

Hence, BP still applies

- Also, zero pruning distances help keeping floating point errors under control

Re-orders

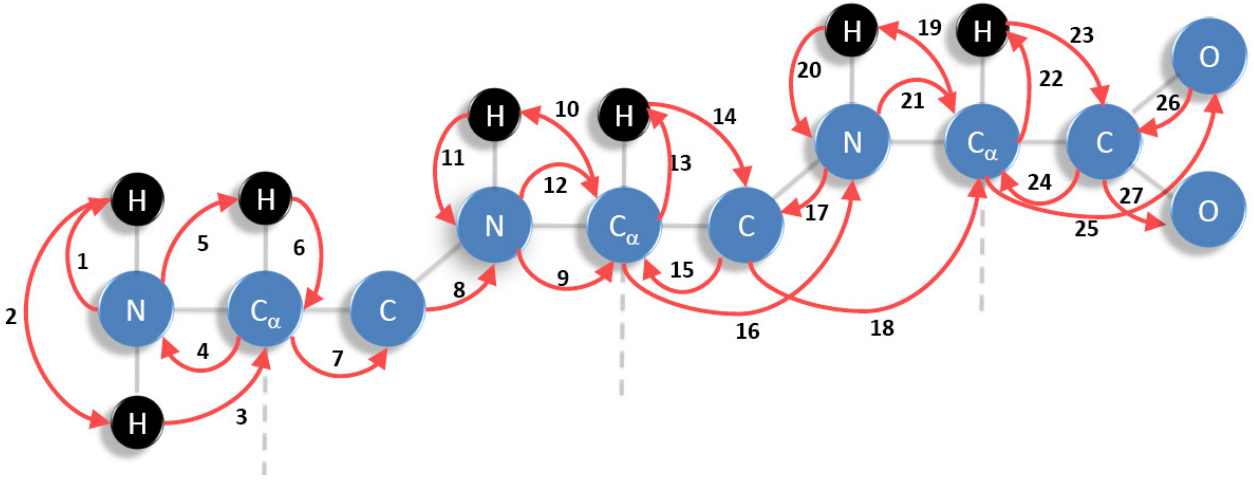
Defn.

A *repetition order* (re-order) is a finite sequence on V

- Re-orders generalize “counting vertices more than once”
- They add more flexibility to exploit certain distances as discretization distances
- Essentially, they provide a tool with which to hand-craft convenient vertex orders for interesting instance classes

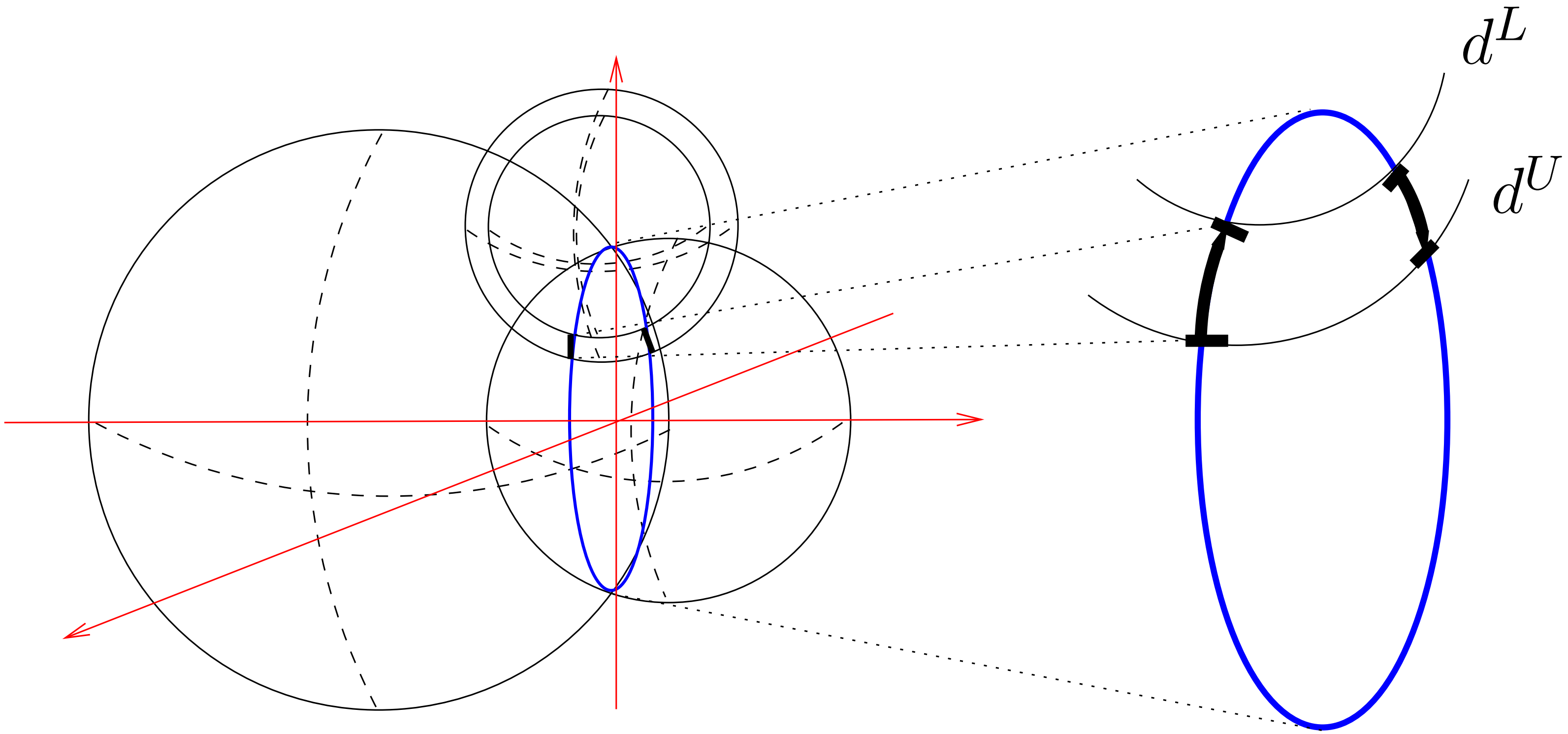
Not immediately evident how to best order proteins

Here’s a re-order applying to all backbones



Uncertain distances

- Typically, NMR provides uncertain distances, modelled by intervals $[d^L_{uv}, d^U_{uv}]$
- Cannot be used for discretization



Two precise distances and an uncertain one

The actual situation

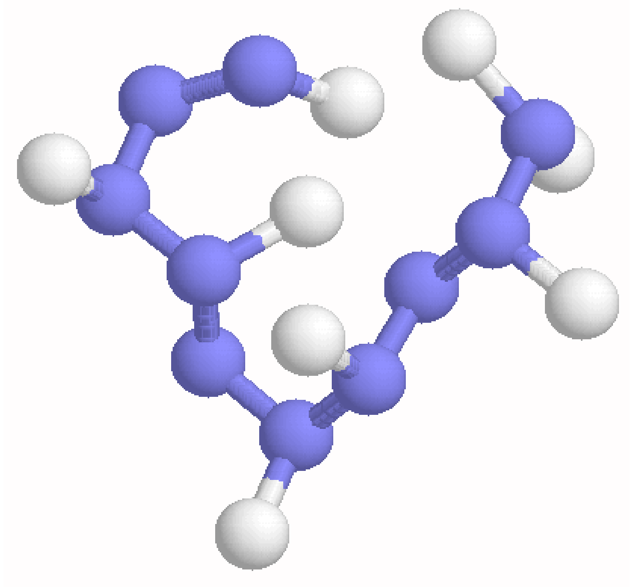
- We know several distances d_{uv} precisely because of chemical properties
- Some distances take values in a finite set D_{uv}
- The distribution of **precise/discrete/uncertain** distances on the protein backbone does not satisfy the DMDGP requirements

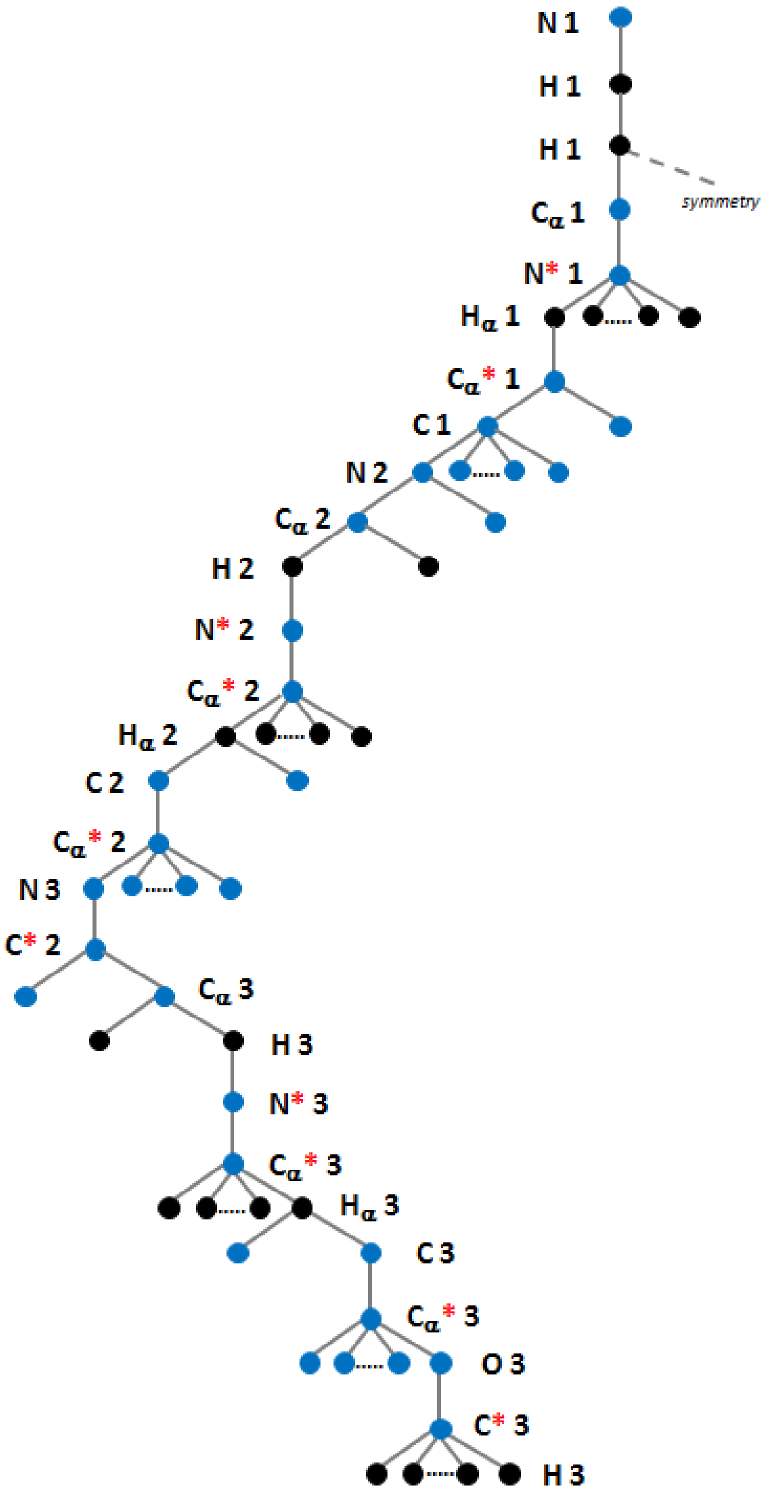
Re-orders provide a solution: use all **precise** distances for discretization, plus a few of the **discrete** whenever needed; **uncertain** distances are used for pruning

- Pruning with intervals is easy: if the current point x_v is s.t. $\|x_v - x_u\| \in [d^L_{uv}, d^U_{uv}]$ for all $u \in \alpha(v)$ accept it, otherwise prune it
- Discrete distances D_{uv} simply give rise to BP nodes at level $v - 1$ with potentially $2|D_{uv}|$ subnodes

*i*BP

[Mucherino et al. SEA11]





Implementations

Sequential code

Mucherino et al. LNCS 2010

- The code is available in open source
- Download: `http://www.antoniomucherino.it/en/mdjeep.php`
- Any doubt, ask the MASTER (Antonio)

Parallel code

Seconds of user CPU on Grid5000 (`www.grid5000.fr`)

	CPUs			
$\|V\|$	1	2	8	64
5000	3.21	1.30	0.54	0.36
7500	4.73	3.15	1.25	0.93
10000	13.38	5.49	2.49	1.57

Embed subgraphs then glue embeddings (rigidity $\Rightarrow$ exact)

A selection of current work

- Work with biochemists/bioinformaticians at Institut Pasteur to access and treat real NMR data
- Use $\mathcal{G}_P x = X$ result from symmetry to obtain all solutions from just one
- Extend complexity study to actual problem with discrete/uncertain distances
- Progress on “MDGP $\in$ **NP**?” question

The end

- **Survey 1**: Liberti, Lavor, Mucherino, Maculan, *Molecular distance geometry methods: from continuous to discrete*, International Transactions in Operational Research, 18:33-51, 2010
- **Survey 2**: Lavor, Liberti, Maculan, Mucherino, *Recent advances on the discretizable molecular distance geometry problem*, European Journal of Operational Research, invited survey (to appear)

Appendix

Continuous formulation

- Solving the system

$$\forall \{i,j\} \in E \quad ||x_i - x_j|| = d_{ij}, \tag{3}$$

is numerically challenging

$\boxed{\text{LHS involves } \sqrt{\text{arg}}\text{, floating point ops} \Rightarrow \text{arg} < 0 \Rightarrow \text{error and abort}}$

$\Rightarrow$ square both sides

- Usually, cast as a penalty objective to be minimized

$$\min_x \sum_{\{i,j\} \in E} (||x_i - x_j||^2 - d_{ij}^2)^2. \tag{4}$$

- Unconstrained minimization of a polynomial of fourth degree

General-purpose methods

- sBB (exact) [L. et al. ’06]: OK on small and medium-sized instances

  because we know the optimal value of the objective (0), lower bound is tight at the initial tree levels

- VNS (heur) [L. et al. ’05, L. et al. ’06]: good for large(ish) instances

- MultiLevel Single Linkage (heur) [Kucherenko et al. ’06]: so-so

		sBB		VNS		MLSL	
Atoms	Variables	OF Value	Time	OF Value	Time	OF Value	Time
cube8	24	0	0.22	0	1.21	0	13.56
cube27	81	0	30.39	0	34.01	0	300.285
cube64	192	0	2237.73	0	398.875	0	2765.13
lavor5	15	0	0.02	0	0.48	0	0.57
lavor10	30	0	1.12	0	7.06	0	69.71
lavor20	60	0	2.25	0	49.99	0	411.152
lavor30	90	0	488.87	0	352.06	0	1634.09
lavor40	120	-	-	0.09	1258.13	0.547	2376.01
lavor50	150	-	-	0	673.48	0	3002.88

MDGP-specific methods

- Smoothing-based:
  - Continuation method (heur) [Moré, Wu ’97]
  - Double VNS with smoothing (heur) [L. et al. ’09]
  - DC optimization with smoothing (heur) [An et al. ’03]
  - Hyperbolic smoothing (heur) [Xavier ’08]
- Alternating projections algorithm (heur) [Glunt et al. 90]:

  iterative updating of a dissimilarity matrix
- Geometric build-up (exact/heur) [Dong, Wu ’03 and ’07]: *triangulation*
- GNOMAD (heur) [Williams et al. ’01]

  iterative updating of atomic ordering minimizing error contribution
- Monotonic Basin Hopping (heur) [Grosso et al. ’09]

  funnel-based population heuristic
- Self-organization heuristic (heur) [Xu et al. ’03]

  pairwise atomic position modification heuristic
- SDP-based formulation *[Ye et al. ’09]*

Geometric build-up

[Dong, Wu ’03], [Dong, Wu ’07]

Given $U = \{1,2,3,4\} \subseteq V$ and a partial embedding $x : U \to \mathbb{R}^3$

1. Consider $v \in V \smallsetminus U$ s.t. $U \subseteq \delta(v)$

2. Extend x to v by solving a linear system:

$$\boxed{\begin{array}{rcl} \|x_v - x_1\|^2 & = & d_{1v}^2 \\ \|x_v - x_2\|^2 & = & d_{2v}^2 \\ \|x_v - x_3\|^2 & = & d_{3v}^2 \\ \|x_v - x_4\|^2 & = & d_{3v}^2 \end{array}} \Rightarrow \boxed{\begin{array}{rcl} \|x_v\|^2 - 2x_v \cdot x_1 + \|x_1\|^2 & = & d_{1v}^2 \text{ (5)} \\ \|x_v\|^2 - 2x_v \cdot x_2 + \|x_2\|^2 & = & d_{1v}^2 \text{ (6)} \\ \|x_v\|^2 - 2x_v \cdot x_3 + \|x_3\|^2 & = & d_{1v}^2 \text{ (7)} \\ \|x_v\|^2 - 2x_v \cdot x_4 + \|x_4\|^2 & = & d_{1v}^2 \text{ (8)} \end{array}}$$

$$\boxed{\begin{array}{l} \text{(8)-(5)} \\ \text{(8)-(6)} \\ \text{(8)-(7)} \end{array}} \Rightarrow \boxed{\begin{pmatrix} 2(x_1 - x_4) \\ 2(x_2 - x_4) \\ 2(x_3 - x_4) \end{pmatrix} x_v = \begin{pmatrix} (\|x_1\|^2 - \|x_4\|^2) - (d_{1v}^2 - d_{4v}^2) \\ (\|x_2\|^2 - \|x_4\|^2) - (d_{2v}^2 - d_{4v}^2) \\ (\|x_3\|^2 - \|x_4\|^2) - (d_{3v}^2 - d_{4v}^2) \end{pmatrix}}$$

3. Let $U \leftarrow U \cup \{v\}$; if $U = V$ stop otherwise repeat from Step 1

Exact on complete and 3-trilateration graphs, heuristic otherwise