Euclidean Distance Geometry

Leo Liberti

IBM Research, USA
CNRS LIX Ecole Polytechnique, France

MFD 2014, Campinas

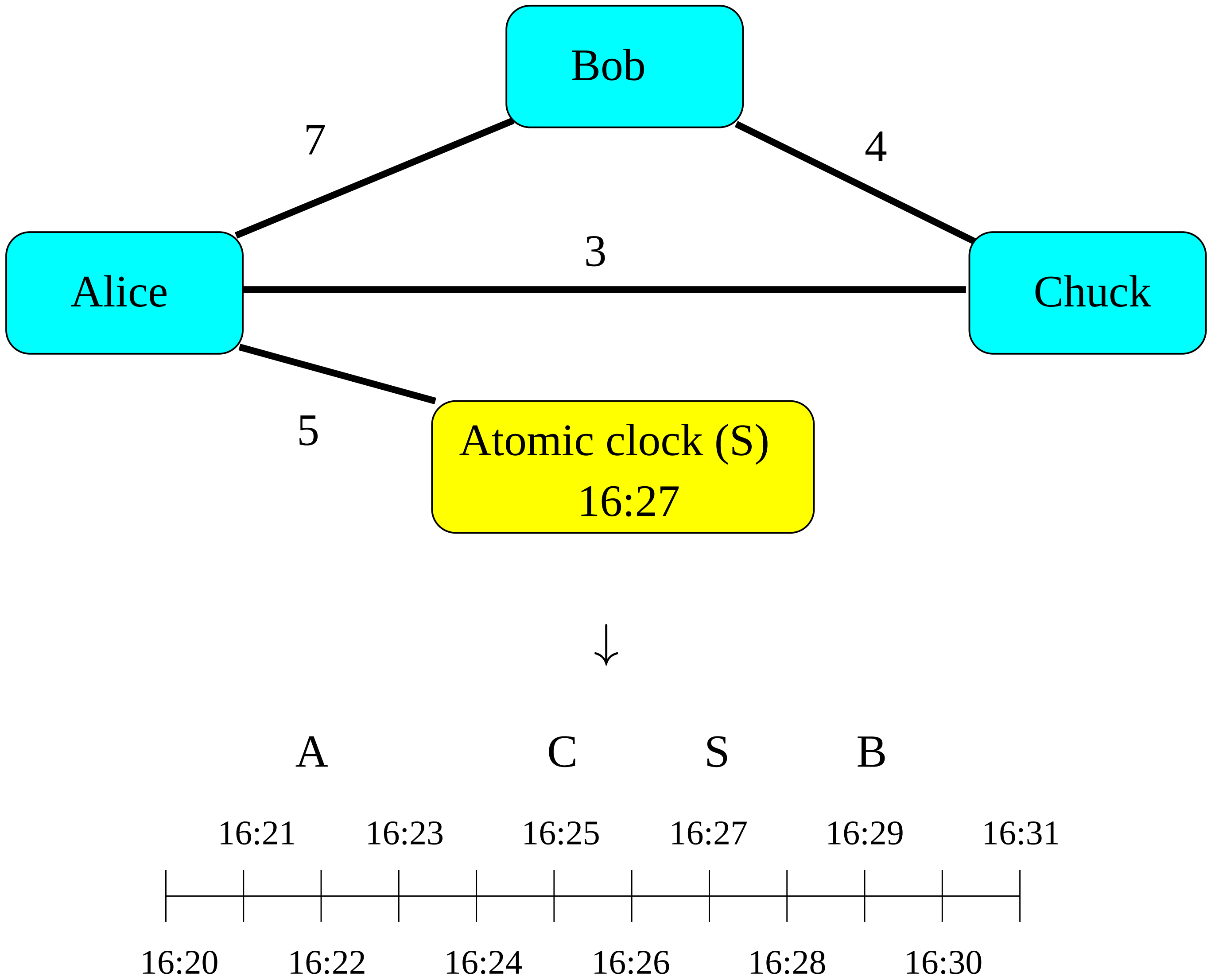
[L., Lavor: *Introduction to Euclidean Distance Geometry*, in preparation]

Table of contents

Applications

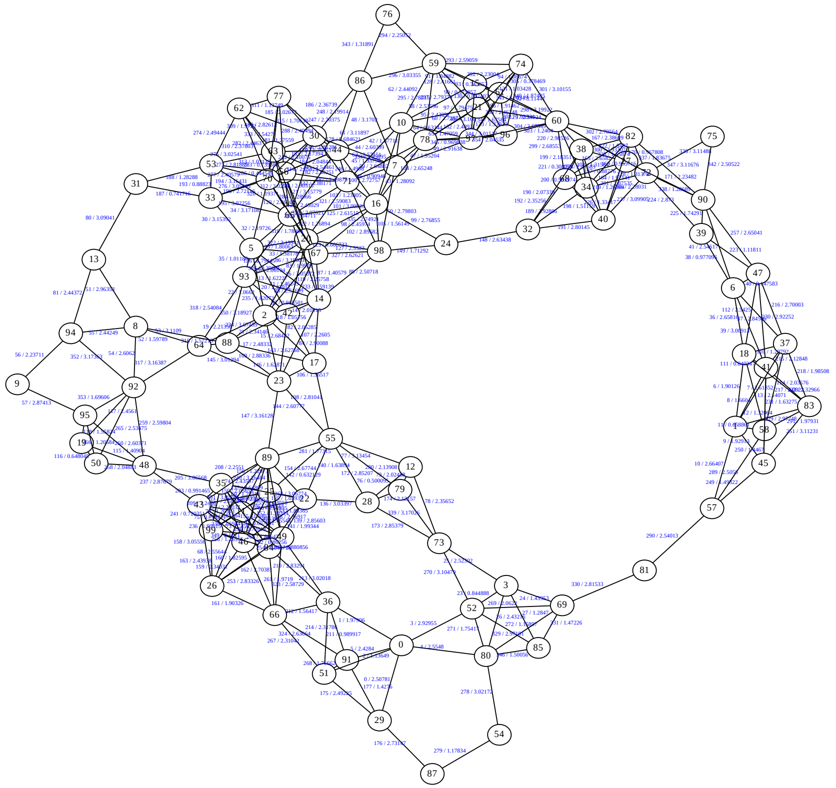
1. **Applications**
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

Clock Synchronization

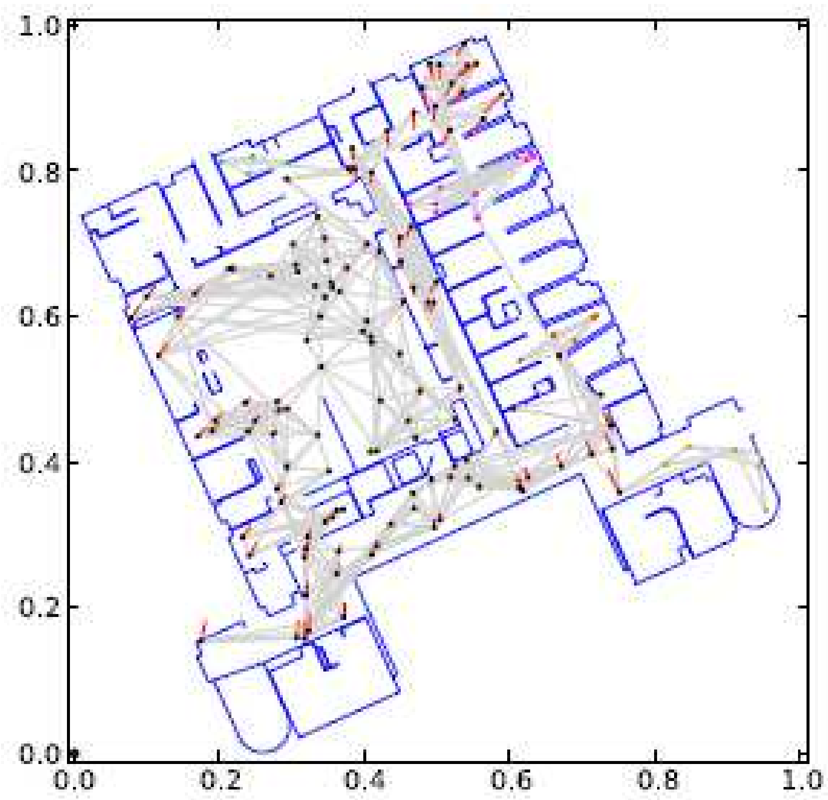


[Singer, 2011]

Sensor network localization

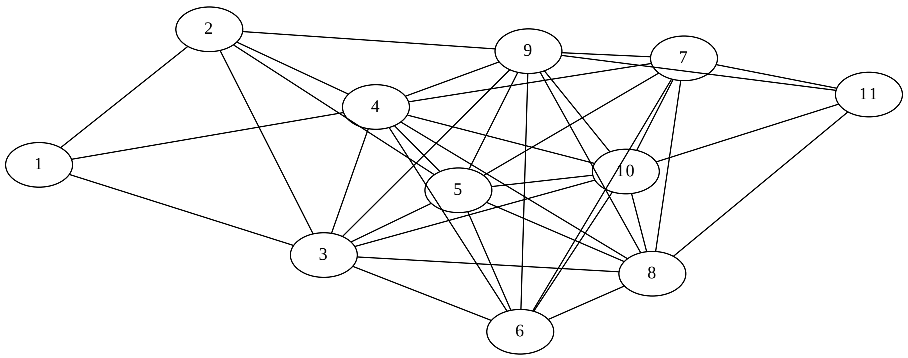


$\longrightarrow$

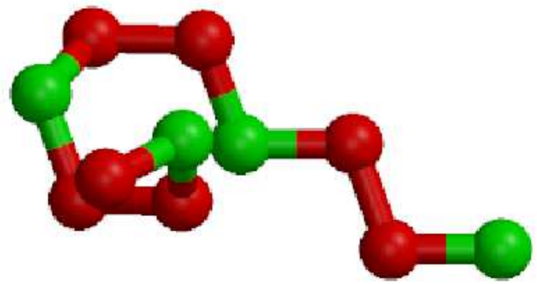


[Yemini, 1978]

Protein conformation from NMR data



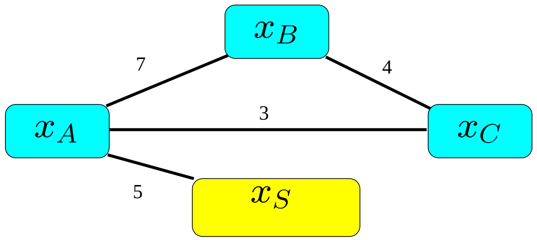
⟶

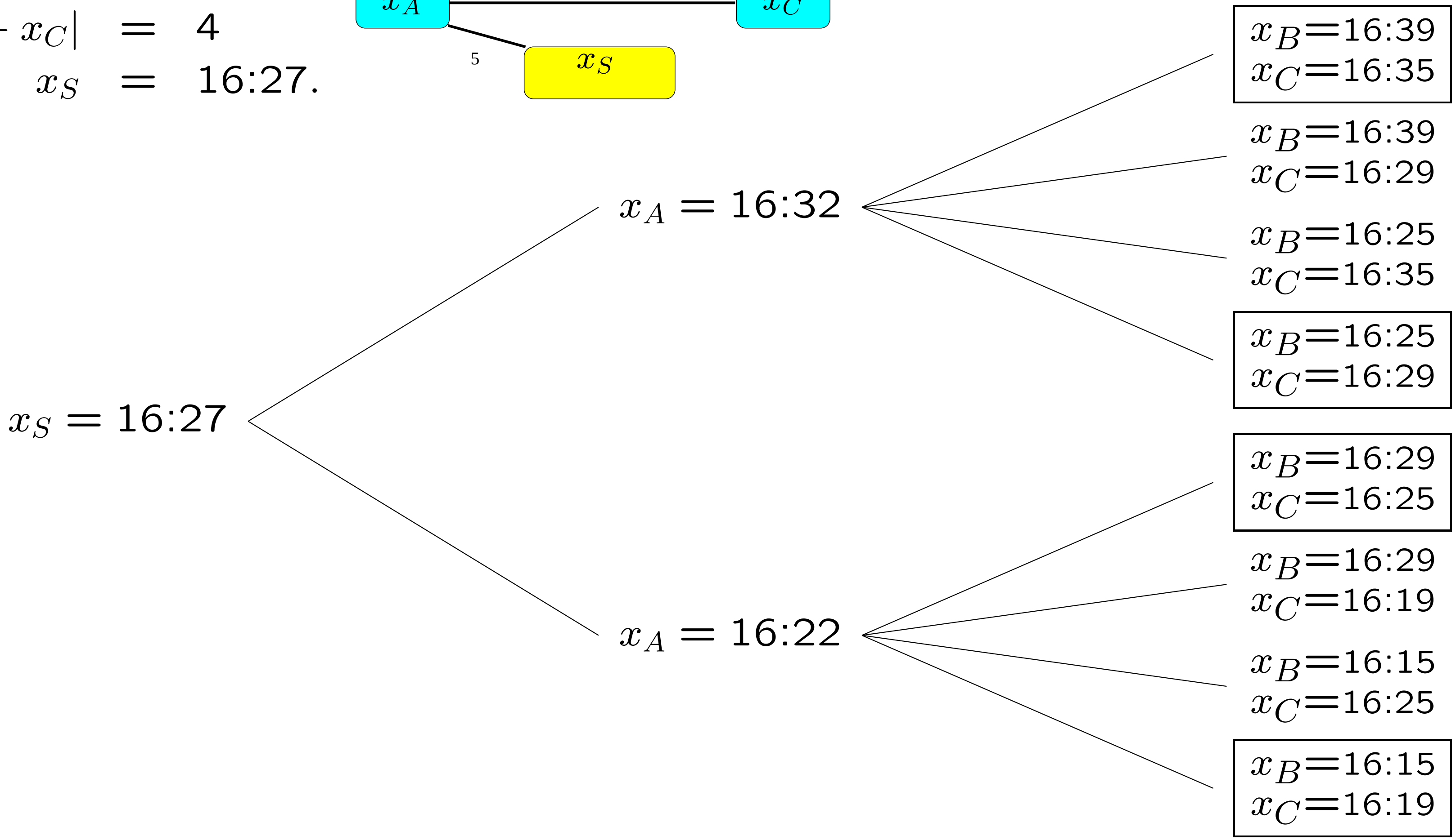


[Crippen & Havel 1988]

Clock synchronization: solutions

$$
\begin{aligned}
|x_A - x_B| &= 7 \\
|x_A - x_C| &= 3 \\
|x_A - x_S| &= 5 \\
|x_B - x_C| &= 4 \\
x_S &= \text{16:27}.
\end{aligned}
$$



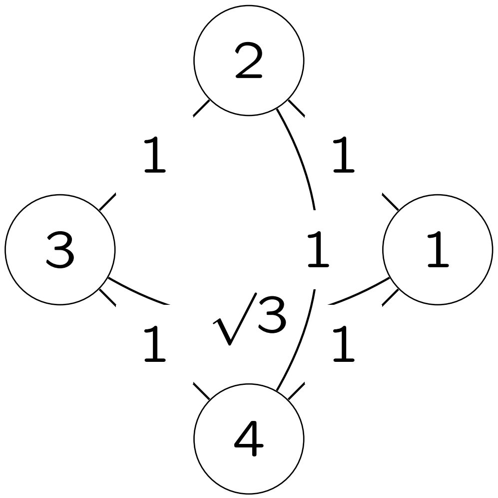


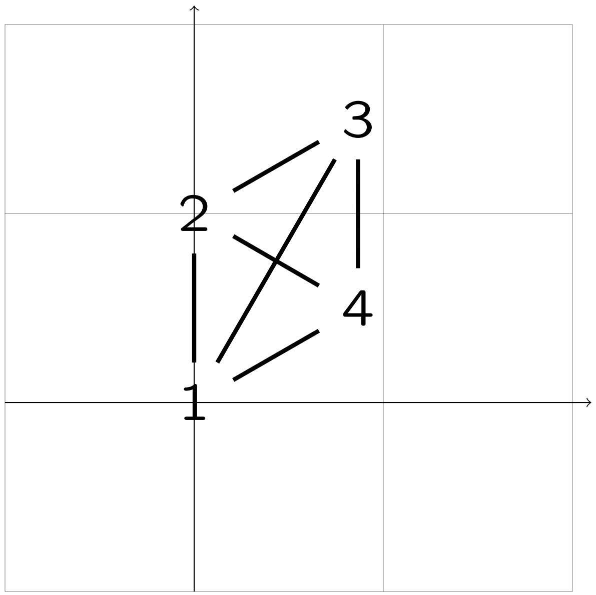
Distance Geometry Problem (DGP)

Given:
– a simple graph $G = (V, E)$
– an edge function $d : E \to \mathbb{R}_{\geq 0}$
– an integer $K \in \mathbb{N}$

Determine whether $\exists$:

a *realization* $x : V \to \mathbb{R}^K$ s.t.
$\forall \{u, v\} \in E \quad \|x_u - x_v\|_2 = d_{uv}$





Let $n = |V|$

More applications

- Autonomous underwater vehicles [Bahr et al. 2009]
- Statics of rigid structures [Maxwell 1864]
- Matrix completion [Laurent 2009]
- Statistics [Boer 2013]
- Psychology [Kruskal 1964]

[Liberti et al., SIREV 2014]

Complexity primer

Definitions

- *Decision problem*: mathematical YES/NO-type question depending on a parameter vector π

- *Instance*: same as above with π replaced by given values v

- *Certificate*: proof that a given answer is true

- **P**: all decision problems solvable in at most $p(|\pi|)$ steps where p is a polynomial

- **NP**: all decision problems with |YES certificate| $\leq p(|\pi|)$ where p is a polynomial

Reductions

- P, Q: decision problems

- If $\exists$ algorithm A which:

  1. reformulates instances $\bar{P}$ of P into instances $\bar{Q}$ of Q
  2. has answer($\bar{P}$) $=$ YES iff answer($A(\bar{Q})$) $=$ YES
  3. is polytime in the *instance size* $|\bar{P}|$

  then A is a *reduction* of P to Q

NP-hardness

- Q is **NP**-hard if every problem in **NP** reduces to Q

- Q is **NP**-complete if it is **NP**-hard and is in **NP**

Why does it work?

any P in **NP** —(polytime reduction)→ Q: how hard?

- Suppose Q easier than P
- Solve P by reducing to Q in polytime and then solve Q
- Then P as easy as Q, against assumption
- $\Rightarrow$ Q at least as hard as P

So if Q is **NP**-hard it is as hard as any problem in **NP**

$\Rightarrow$ Q **is as hard as the hardest problem in NP**

NP-hardness proofs

Given a new problem Q, take any known **NP**-hard problem P and reduce it to Q

Why does it work?

P: **NP**-hard $\xrightarrow{\text{polytime reduction}}$ Q: how hard?

- **As before**: Suppose . . . (etc.) $\Rightarrow$ Q at least as hard as P
- Since P is **NP**-hard, it is hardest in **NP**, and so is Q

$\Rightarrow$ Q is **NP**-hard

Complexity of the DGP

1. Applications
2. Definition
3. Complexity primer
4. **Complexity of the DGP**
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

DGP $\in$ NP?

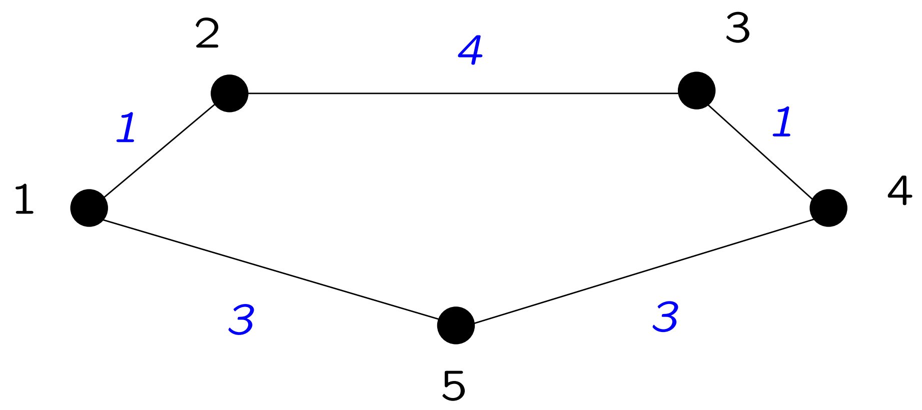
- **NP**: YES/NO problems with polytime-checkable proofs for YES
- DGP is a YES/NO problem
- DGP$_1$ $\in$ **NP**, since $d_{uv} = |x_u - x_v| \Rightarrow (d \in \mathbb{Q} \to x \in \mathbb{Q})$
- Solutions might involve irrational numbers when $K > 1$
- Some empirical evidence that DGP $\notin$ **NP** [Beeker et al. 2013]

The DGP is NP-hard

Partition

Given $a = (a_1, \ldots, a_n) \in \mathbb{N}^n$, $\exists\, I \subseteq \{1, \ldots, n\}$ s.t. $\sum_{i \in I} a_i = \sum_{i \notin I} a_i$?

- Reduce (**NP**-hard) Partition to DGP$_1$
- $a \longrightarrow$ cycle C with $V(C) = \{1, \ldots, n\}$, $E(C) = \{\{1,2\}, \ldots, \{n,1\}\}$
- For $i < n$ let $d_{i,i+1} = a_i$, and $d_{n,n+1} = d_{n1} = a_n$
- E.g. for $a = (1,4,1,3,3)$, get cycle graph:



[Saxe, 1979]

Partition is YES $\Rightarrow$ DGP$_1$ is YES

- **Given**: $I \subset \{1, \ldots, n\}$ s.t. $\sum_{i \in I} a_i = \sum_{i \notin I} a_i$

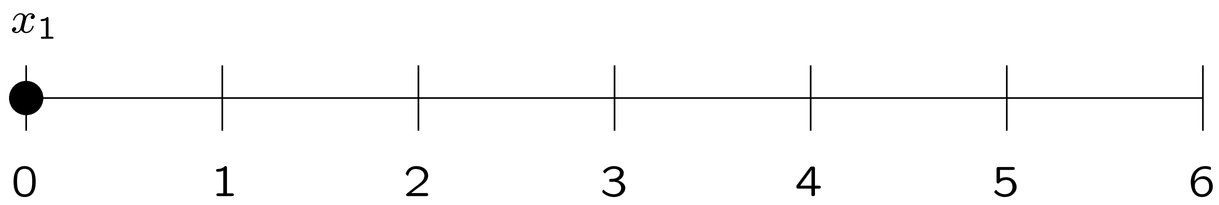
- **Construct**: realization x of C in $\mathbb{R}$

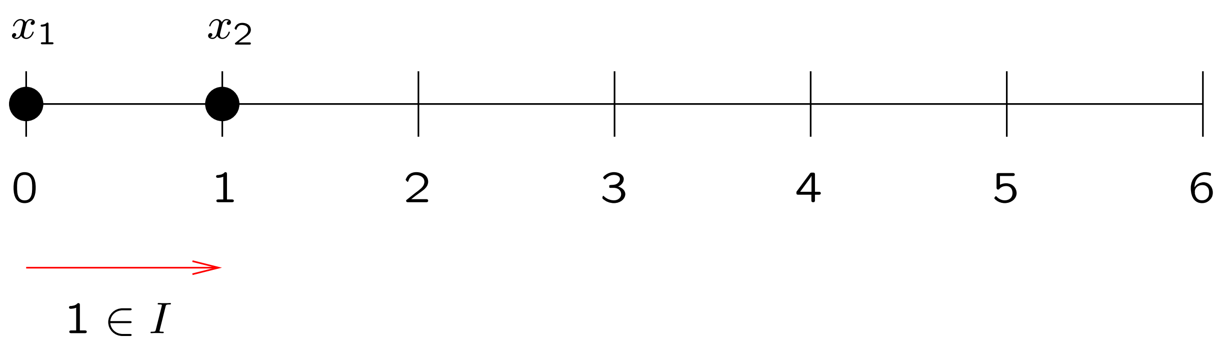
  1. $x_1 = 0$ `// start`

  2. **induction step**: suppose x_i known
     if $i \in I$
     let $x_{i+1} = x_i + d_{i,i+1}$ `// go right`
     else
     $x_{i+1} = x_i - d_{i,i+1}$ `// go left`

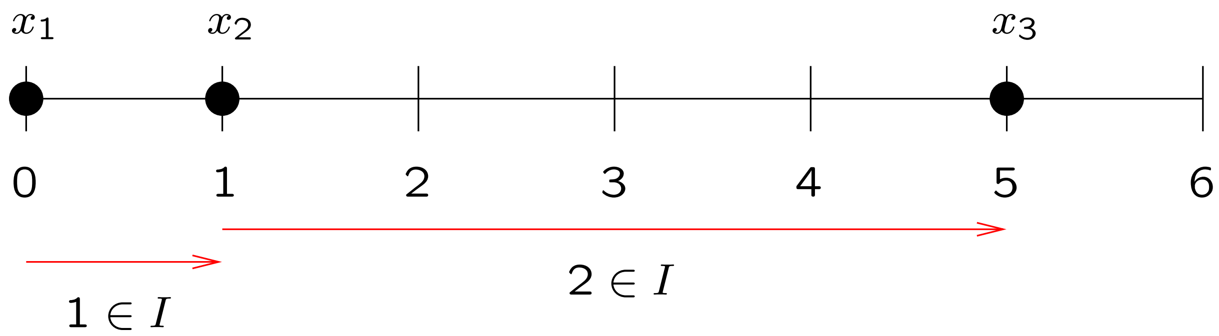
- **Correctness proof**: by the same induction
  but careful when $i = n$: have to show $x_{n+1} = x_1$

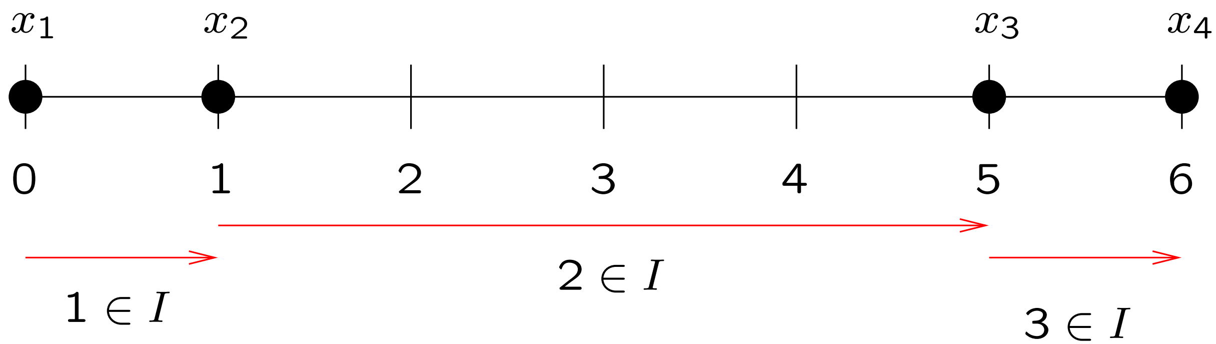
$$I = \{1, 2, 3\}$$

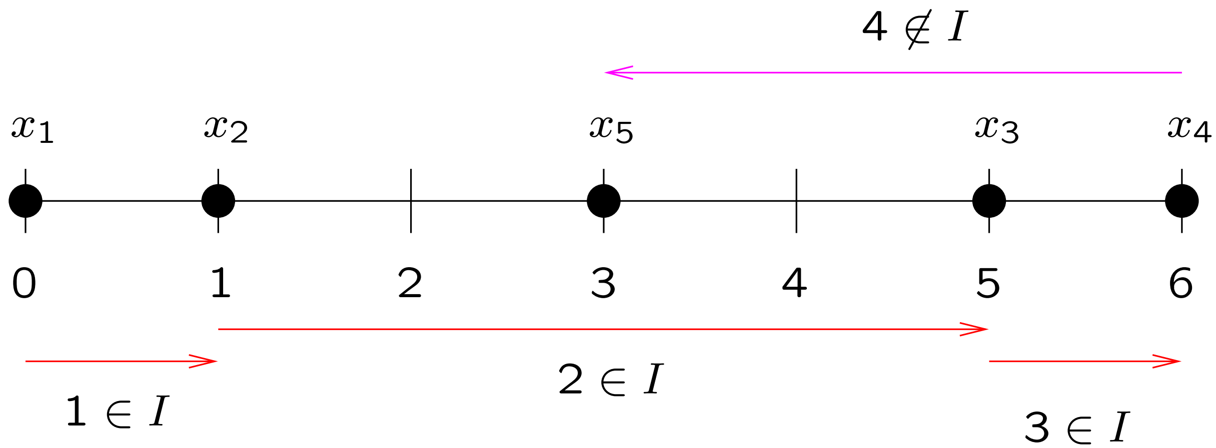




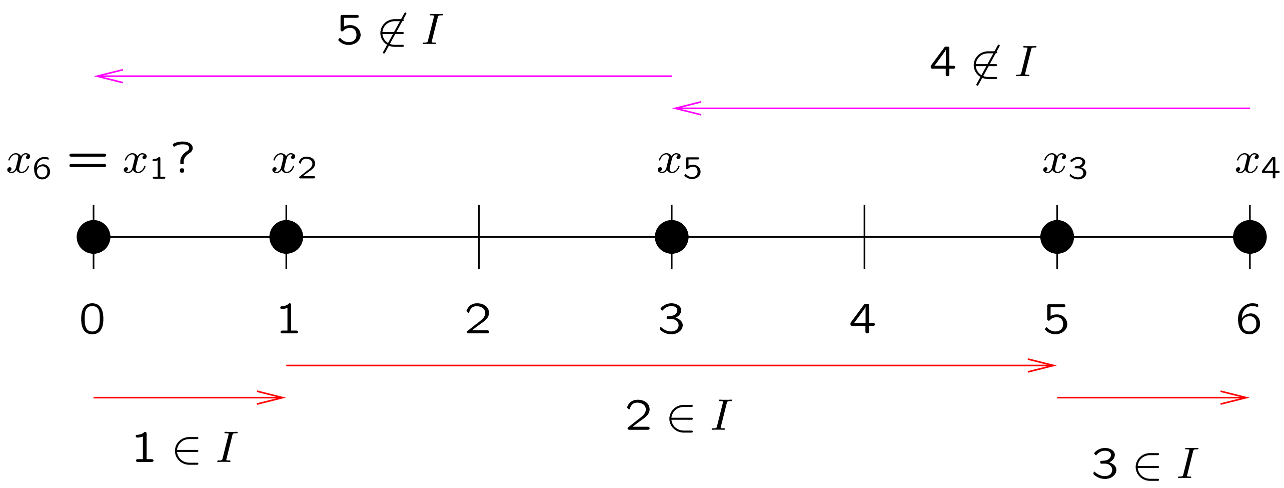
x_1
x_2
0
1
2
3
4
5
6
$1 \in I$


x_1
x_2
x_3
0
1
2
3
4
5
6
$2 \in I$
$1 \in I$


x_1
x_2
x_3
x_4
0
1
2
3
4
5
6
$2 \in I$
$1 \in I$
$3 \in I$



$4 \notin I$
x_1
x_2
x_5
x_3
x_4
0
1
2
3
4
5
6
$2 \in I$
$1 \in I$
$3 \in I$



$5 \notin I$
$4 \notin I$
$x_6 = x_1$?
x_2
x_5
x_3
x_4
0
1
2
3
4
5
6
$2 \in I$
$1 \in I$
$3 \in I$

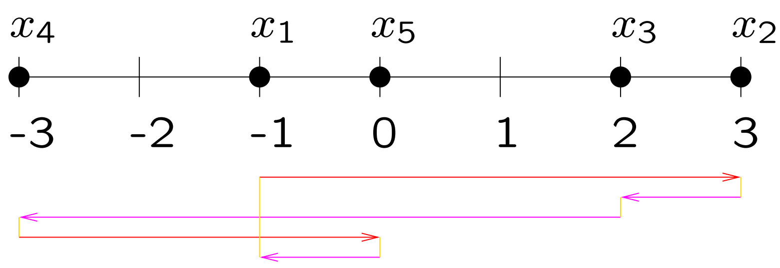
Partition is YES $\Rightarrow$ DGP$_1$ is YES

$$
\begin{aligned}
(1) = \sum_{i\in I}(x_{i+1}-x_i) &= \sum_{i\in I} d_{i,i+1} = \\
= \sum_{i\in I} a_i &= \sum_{i\notin I} a_i = \\
= \sum_{i\notin I} d_{i,i+1} &= \sum_{i\notin I}(x_i - x_{i+1}) = (2)
\end{aligned}
$$

$$
\begin{aligned}
(1) = (2) \Rightarrow \sum_{i\in I}(x_{i+1}-x_i) = \sum_{i\notin I}(x_i - x_{i+1}) \Rightarrow \sum_{i\leq n}(x_{i+1}-x_i) &= 0 \\
\Rightarrow (x_{n+1}-x_n) + (x_n - x_{n-1}) + \cdots + (x_3 - x_2) + (x_2 - x_1) &= 0 \\
\Rightarrow x_{n+1} &= x_1
\end{aligned}
$$

Partition is NO $\Rightarrow$ DGP$_1$ is NO

- By contradiction: suppose DGP$_1$ is YES, x realization of C
- $F = \{\{u,v\} \in E(C) \mid x_u \le x_v\}$, $E(C) \smallsetminus F = \{\{u,v\} \in E(C) \mid x_u > x_v\}$
- Trace $x_1, \ldots, x_n$: follow edges in F ($\to$) and in $E(C) \smallsetminus F$ ($\leftarrow$)



$$\sum_{\{u,v\}\in F} (x_v - x_u) = \sum_{\{u,v\}\notin F} (x_u - x_v)$$
$$\sum_{\{u,v\}\in F} |x_u - x_v| = \sum_{\{u,v\}\notin F} |x_u - x_v|$$
$$\sum_{\{u,v\}\in F} d_{uv} = \sum_{\{u,v\}\notin F} d_{uv}$$

- Let $J = \{i < n \mid \{i, i+1\} \in F\} \cup \{n \mid \{n,1\} \in F\}$

$$\Rightarrow \quad \sum_{i\in J} a_i = \sum_{i\notin J} a_i$$

- So J solves Partition instance, contradiction
- $\Rightarrow$ DGP is **NP**-hard, DGP$_1$ is **NP**-complete

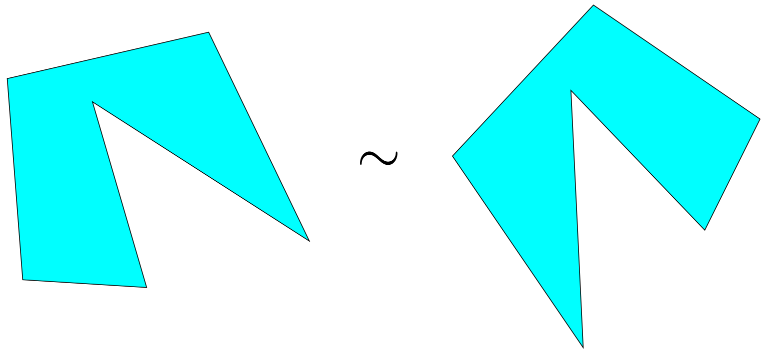
Number of solutions

With congruences

- (G, K): DGP instance
- $\tilde{X} \subseteq \mathbb{R}^{Kn}$: set of solutions
- *Congruence*: composition of translations, rotations, reflections
- $C =$ set of congruences in $\mathbb{R}^K$
- $x \sim y$ means $\exists \rho \in C \ (y = \rho x)$:
  distances in x are preserved in y through ρ
- $\Rightarrow$ if $|\tilde{X}| > 0$, $|\tilde{X}| = 2^{\aleph_0}$

Modulo congruences

- Congruence is an *equivalence relation* $\sim$ on $\tilde{X}$ (reflexive, symmetric, transitive)



- Partitions $\tilde{X}$ into *equivalence classes*

- $X = \tilde{X}/\sim$: sets of representatives of equivalence classes

- **Focus on $|X|$ rather than $|\tilde{X}|$**

Cardinality of X

- infeasible $\Leftrightarrow |X| = 0$
- rigid graph $\Leftrightarrow |X| < \aleph_0$
- globally rigid graph $\Leftrightarrow |X| = 1$
- flexible graph $\Leftrightarrow |X| = 2^{\aleph_0}$
- $|X| = \aleph_0$: impossible by Milnor's theorem

Milnor's theorem implies $|X| \neq \aleph_0$

- System S of polynomial equations of degree d

$$\forall i \leq m \quad p_i(x_1, \ldots, x_{nK}) = 0$$

- Let X be the set of $x \in \mathbb{R}^{nK}$ satisfying S

- **Number of connected components of X is $\leq d(2d-1)^{nK-1}$** [Milnor 1964]

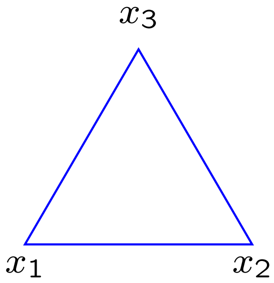
- If $|X|$ is countable then G cannot be flexible
$\Rightarrow$ incongruent elements of X are separate connected components
$\Rightarrow$ by Milnor's theorem, there's finitely many of them

Examples

$V^1 = \{1, 2, 3\}$

$E^1 = \{\{u, v\} \mid u < v\}$

$d^1 = 1$



ρ congruence in $\mathbb{R}^2$

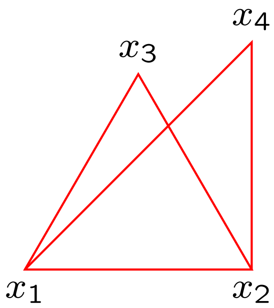
$\Rightarrow \rho x$ valid realization

$|X| = 1$

$V^2 = V^1 \cup \{4\}$

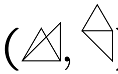
$E^2 = E^1 \cup \{\{1, 4\}, \{2, 4\}\}$

$d^2 = 1 \wedge d_{14} = \sqrt{2}$



ρ reflects x_4 wrt $\overline{x_1, x_2}$

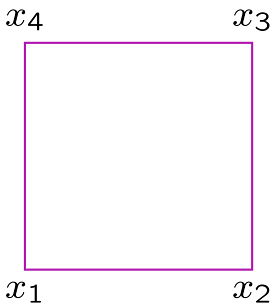
$\Rightarrow \rho x$ valid realization

$|X| = 2$ ()

$V^3 = V^2$

$E^3 = \{\{u, u + 1\} | u \leq 3\} \cup \{1, 4\}$

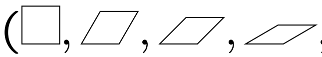
$d^1 = 1$



ρ rotates $\overline{x_2 x_3}$, $\overline{x_1 x_4}$ by θ

$\Rightarrow \rho x$ valid realization

$|X|$ is uncountable

( $, \ldots$)

Mathematical optimization formulations

1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. **Mathematical optimization formulations**
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

System of quadratic constraints

$$\forall \{u,v\} \in E \qquad \|x_u - x_v\|^2 = d_{uv}^2$$

- Around 10 vertices
- Computationally useless

Quadratic objective

$$\min_{x \in \mathbb{R}^{nK}} \sum_{\{u,v\} \in E} (\|x_u - x_v\|^2 - d_{uv}^2)^2$$

- Globally optimal value **zero** iff x is a realization of G
- sBB: 10-100 vertices, exact solutions
- heuristics: 100-1000 vertices, poor quality

[Lavor et al., 2006]

Convexity and concavity

$$\max_{x \in \mathbb{R}^{nK}} \quad \sum_{\{u,v\} \in E} \|x_u - x_v\|^2$$
$$\forall \{u,v\} \in E \quad \|x_u - x_v\|^2 \le d_{uv}^2$$

- Convex constraints, concave objective
- Computationally no better than "quadratic objective"

Pointwise reformulation

$$
\begin{array}{rl}
\max\limits_{x\in\mathbb{R}^{nK}} & \sum\limits_{\{u,v\}\in E, k\le K} \theta_{uvk}(x_{uk}-x_{vk}) \\
\forall \{u,v\}\in E & \|x_u - x_v\|^2 \le d_{uv}^2
\end{array}
$$

- Convex subproblem in stochastic iterative heuristics
  "guess θ and solve"
- 100-1000 vertices, good quality

[L. IOS14/MAGO14(slides)]

SDP formulation

$$\min_{X \succeq 0} \sum_{\{u,v\} \in E} (X_{uu} + X_{vv} - 2X_{uv})$$

$$\forall \{u,v\} \in E \quad X_{uu} + X_{vv} - 2X_{uv} \geq d_{uv}^2$$

- Similar to those of Ye, Wolkowicz — works better for proteins
- 100 vertices, good quality

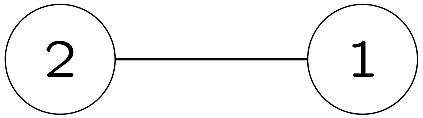
[D'Ambrosio et al., in progress]

Realizing complete graphs

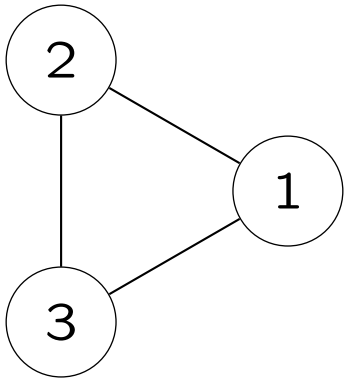
1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. **Realizing complete graphs**
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

Cliques

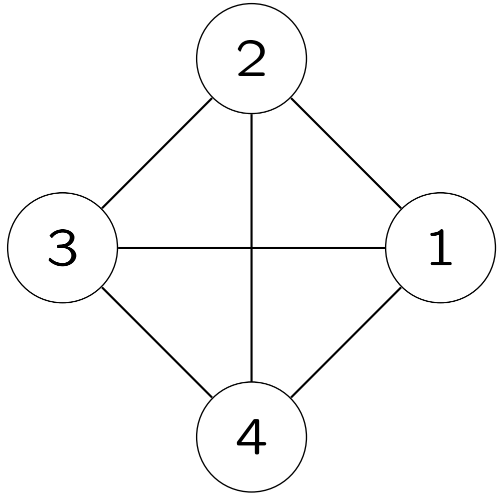
2-clique



3-clique

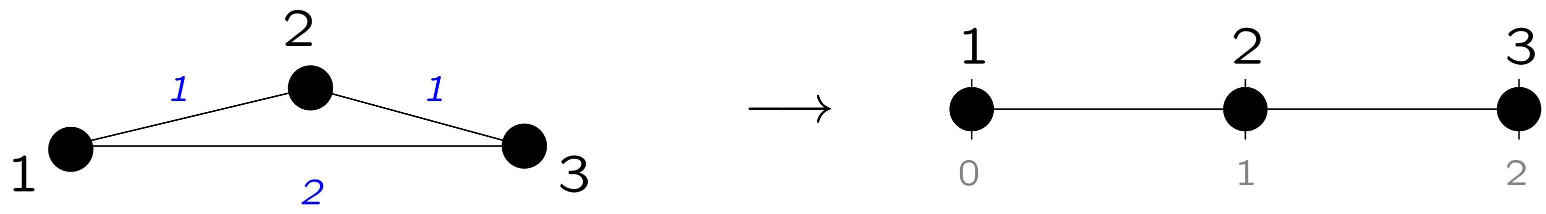


4-clique



$(K+1)$-clique $=$ K-clique $\oplus$ a vertex

Triangulation



Example: realize triangle on a line

- From $\|x_3 - x_1\| = 2$ and $\|x_3 - x_2\| = 1$ get

$$x_3^2 - 2x_1x_3 + x_1^2 = 4 \quad (1)$$
$$x_3^2 - 2x_2x_3 + x_2^2 = 1. \quad (2)$$

- (**??**) – (**??**) yields

$$2x_3(x_1 - x_2) = x_1^2 - x_2^2 - 3$$
$$\Rightarrow \quad 2x_3 = 4,$$

- Hence $x_3 = 2$

Realizing a $(K+1)$-clique in $\mathbb{R}^{K-1}$

- Apply triangulation inductively on K
  assume $x_1,\ldots,x_K \in \mathbb{R}^{K-1}$ known, compute $y = x_{K+1}$

- K quadratic eqns ($\forall j \le K \; \|y - x_j\|^2 = d^2_{j,K+1}$) in $K-1$ vars

$$\left\{\begin{array}{rcll} \|y\|^2 - 2x_1\cdot y + \|x_1\|^2 & = & d^2_{1,K+1} & [1] \\ & \vdots & & \vdots \\ \|y\|^2 - 2x_K\cdot y + \|x_K\|^2 & = & d^2_{K,K+1} & [K] \end{array}\right.$$

- Form system $\forall j \le K$ ($[j]-[K]$)

$$\left\{\begin{array}{rcll} 2(x_1 - x_K)\cdot y & = & \|x_1\|^2 - \|x_K\|^2 - d^2_{1,K+1} + d^2_{K,K+1} & [1]-[K] \\ & \vdots & & \vdots \\ 2(x_{K-1} - x_K)\cdot y & = & \|x_{K-1}\|^2 - \|x_K\|^2 - d^2_{K-1,K+1} + d^2_{K,K+1} & [K-1]-[K] \end{array}\right.$$

- This is a $(K-1)\times(K-1)$ linear system $Ay = b$

Solve to find y

[Dong, Wu 2002]

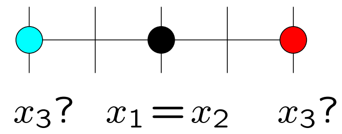
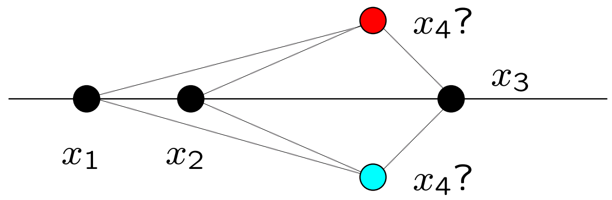
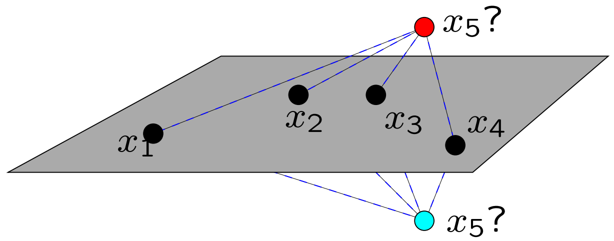
“Solve”?

1. What if A is singular?

2. Or: A nonsingular but instance is NO

Singularity: rk$A = K - 2$

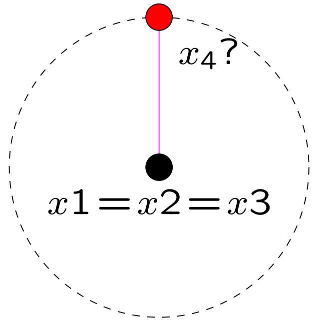
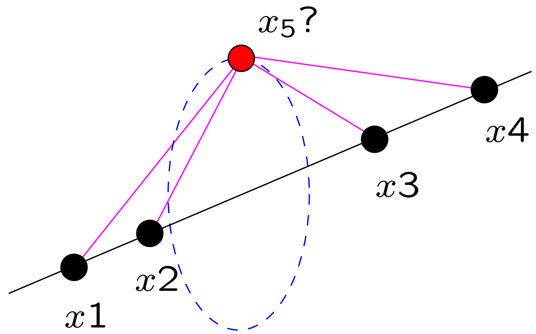
One row $x_j - x_K$ of A depends on the others

$K = 2$	triangle in $\mathbb{R}^1$	$x_1 - x_2 = 0$	
$K = 3$	4-clique in $\mathbb{R}^2$	x_1, x_2, x_3 on a line	
$K = 4$	5-clique in $\mathbb{R}^3$	$x_1, \ldots, x_4$ in a plane	

Trend continues: rk $A = K - 2 \Rightarrow |X| = 2$ (see later)

Singularity: rk$A = K - 3$

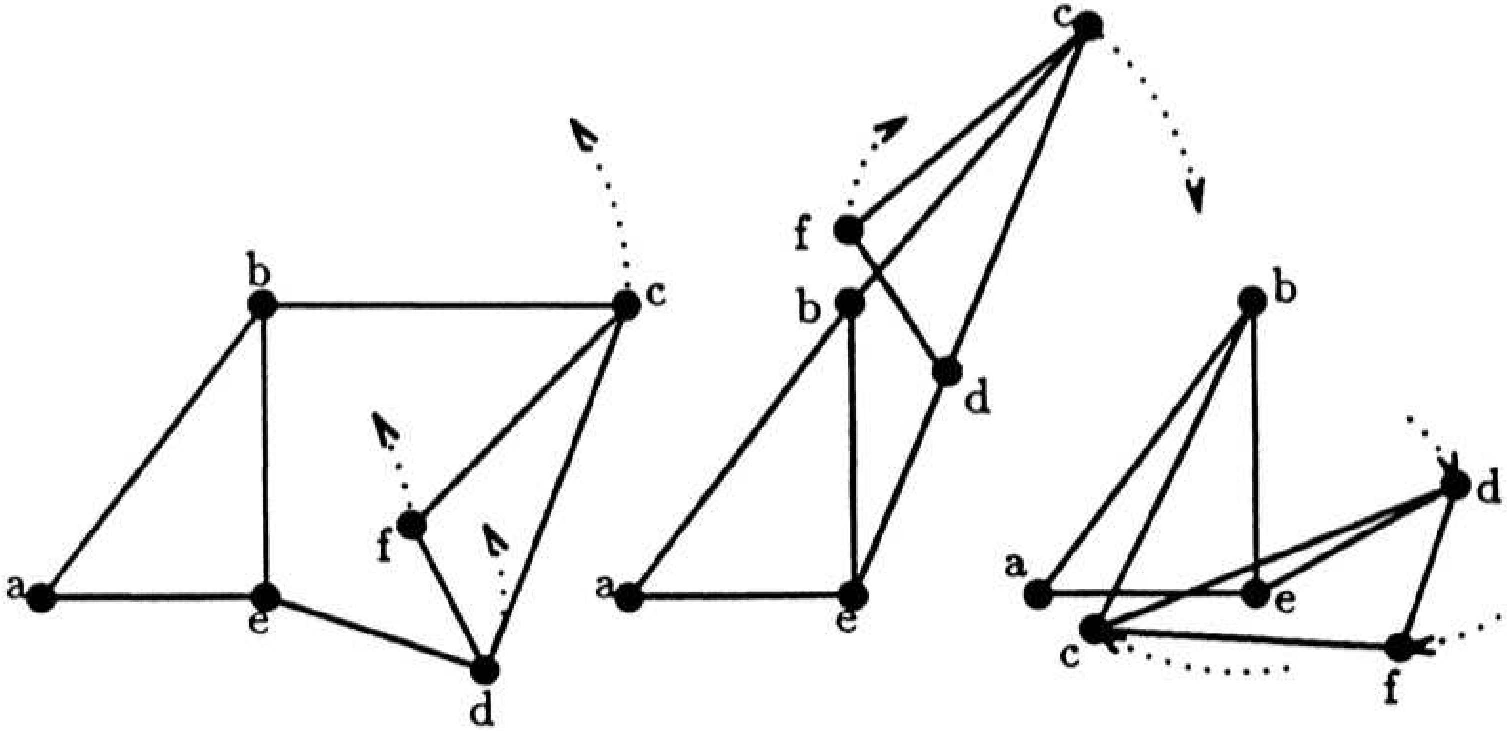
Two rows $x_j - x_k$ depend on the others

$K = 3$	4-clique in $\mathbb{R}^2$	$x_1 = x_2 = x_3$	
$K = 4$	5-clique in $\mathbb{R}^3$	$x_1, \ldots, x_4$ on a line	

Trend continues: [Hendrickson, 1992]

Thm. 5.8. *If a graph G is connected, flexible and has more than K vertices, $|X|$ contains almost always a submanifold diffeomorphic to a circle*

Hendrickson's theorem also applies to non-cliques



Nonsingular matrix A with NO instance

- Infeasible quadratic system $\forall j \leq K+1 \ \|x_j - x_K\|^2 = d_{jK}$
- Take differences, get nonsingular A and value for x_K
- . . . but it's wrong!

Shit happens!

Every time you solve the linear system $Ay = b$
check feasibility with quadratic system

Algorithm for realizing complete graphs in $\mathbb{R}^K$

- Assume:

  (i) $G = (V, E)$ complete

  (ii) $|V| = n \geq K + 2$

  (iii) we know $x_1, \ldots, x_{K+1}$

- Increase K: we know how to realize x_{K+2} in $\mathbb{R}^K$

- Use this inductively for each $i \in \{K + 2, \ldots, n\}$

Algorithm for realizing complete graphs in $\mathbb{R}^K$

```
// realize next vertex iteratively
for i ∈ {K+2, ..., n} do
    // use (K+1) immediate adjacent predecessors to compute x_i
    if rk A = K then
        x_i = A^{-1} b // A, b defined as above
    else
        x_i = ∞ // A singular, mark ∞ and exit
        break
    end if
    // check that x_i is feasible w.r.t. other distances
    for {j ∈ N(i) | j < i} do
        if ||x_i − x_j|| ≠ d_ij then
            // if not, mark infeasible and exit loop *
            x_i = ∅
            break
        end if
    end for
    if x_i = ∅ then
        break
    end if
end for
return x
```

* the “ignore trouble” policy, a.k.a. “ignore probability zero events”

Complexity of Alg. 1

- Outer loop: $O(n)$
- Rank and inverse of A: $O(K^3)$
- Inner loop: $O(n)$
- Get $O(n^2 K^3)$
- But in most applications K is fixed
- **Get** $O(n^2)$

But how do we find the realization of the first $K+1$ vertices?

Realizing $(K+1)$-cliques in $\mathbb{R}^K$

- Realizing $(K+1)$-cliques in $\mathbb{R}^{K-1}$ yields "flat simplices" (e.g. triangles on lines)

- Use "natural" embedding dimension $\mathbb{R}^K$

- Same reasoning as above:
  get system $Ay = b$ where $y = x_{K+1}$ and $A_j = 2(x_j - x_K)$

- **But now A is $(K-1) \times K$**

- *Same as previous case with A singular*

Almost square

How can you solve the following system $Ay = b$:

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{K-1,1} & a_{K-1,2} & \cdots & a_{K-1,K} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_{K-1} \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_{K-1} \end{pmatrix}$$

where A has one more columns than rows and rank $K-1$?

Basics and nonbasics

- Since $\text{rk}\, A = K - 1$, $\exists$ $K - 1$ linearly independent columns
- $\mathcal{B}$: set of their indices
- $\mathcal{N}$: index of remaining columns
- B: $(K - 1) \times (K - 1)$ square matrix of columns in $\mathcal{B}$
- $\Rightarrow$ B is nonsingular
- **Can partition columns as** $A = (B|N)$

  Column j corresponds to variable y_j
- Variables $y_{\mathcal{B}}$ are called *basic variables*
- Variable $y_{\mathcal{N}}$ is called *nonbasic variable*

The dictionary

$$\begin{aligned} (B|N)y &= b \\ \Rightarrow \quad By_{\mathcal{B}} + Ny_{\mathcal{N}} &= b \\ \Rightarrow \quad y_{\mathcal{B}} &= B^{-1}b - B^{-1}Ny_{\mathcal{N}} \end{aligned}$$

Basics expressed in function of nonbasic

One quadratic equation

- From value of $y_{\mathcal{N}}$, can use dictionary to get y
- Use one quadratic equation

1. Pick any $h \in \{1, \ldots, K-1\}$, equation is $\|x_h - y\|_2^2 = d_{hK}^2$
2. $y = (y_{\mathcal{B}} | y_{\mathcal{N}})^\top$
3. Replace $y_{\mathcal{B}}$ with $B^{-1}b - B^{-1}Ny_{\mathcal{N}}$ in equation
4. Solve resulting quadratic equation in one variable $y_{\mathcal{N}}$
5. **Get 0,1 or 2 values for $y_{\mathcal{N}}$**
6. $\Rightarrow$ Get 0,1 or 2 positions for x_{K+1}

What if $B^{-1}N$ is zero?

- $y_{\mathcal{B}} = B^{-1}b - B^{-1}Ny_{\mathcal{N}}$ reduces to $y_{\mathcal{B}} = B^{-1}b$

- Use one quadratic equation

1. Pick any $h \in \{1, \ldots, K-1\}$, equation is $\|x_h - y\|_2^2 = d_{hK}^2$
2. $y = (y_{\mathcal{B}}|y_{\mathcal{N}})^\top$
3. Replace $y_{\mathcal{B}}$ with $B^{-1}b$ in equation
4. Solve resulting quadratic equation in one variable $y_{\mathcal{N}}$
5. **Get 0,1 or 2 values for $y_{\mathcal{N}}$**
6. $\Rightarrow$ Get 0,1 or 2 positions for x_{K+1}

The difference

- $B^{-1}N \neq 0$: $y_{\mathcal{N}} \xrightarrow{\text{dictionary}} y_{\mathcal{B}}$

- Different values $y_{\mathcal{N}}^{+} \neq y_{\mathcal{N}}^{-} \longrightarrow y^{+}, y^{-}$ with different components

- $B^{-1}N = 0$: $y_{\mathcal{B}} \xrightarrow{\text{quadratic eqn.}} y_{\mathcal{N}}$

- Even if $y_{\mathcal{N}}^{+} \neq y_{\mathcal{N}}^{-}$, $K-1$ components of y^{+}, y^{-} are equal
  $\text{aff}(x_1, \ldots, x_{K-1}) = \{y \in \mathbb{R}^K \mid y_{\mathcal{N}} = 0\}$

The case of no solutions

- No realizations exist for this $(K+1)$-clique in $\mathbb{R}^K$

- **DGP instance is NO**

The case of one solution

- Assume for simplicity: $\mathcal{N} = K$, $h = 1$, $B^{-1}N \neq 0$
  Then $\|x_h - y\|^2 = d^2_{h,K+1}$ becomes:

$$\begin{aligned}
\lambda y_K^2 - 2\mu y_K + \nu &= 0, \quad \text{where} \\
\lambda &= \sum_{\ell,j<K} 1 + \beta_{\ell j}^2 a_{jK}^2 \\
\mu &= x_{1K} + \sum_{\ell,j<K} \beta_{\ell j} a_{jK} (\beta_{\ell j} b_\ell - x_{1\ell}) \\
\nu &= \sum_{\ell,j<K} \beta_{\ell j} b_\ell (\beta_{\ell j} b_\ell - 2x_{1\ell}) + \|x_1\|^2 - d^2_{1,K+1}
\end{aligned}$$

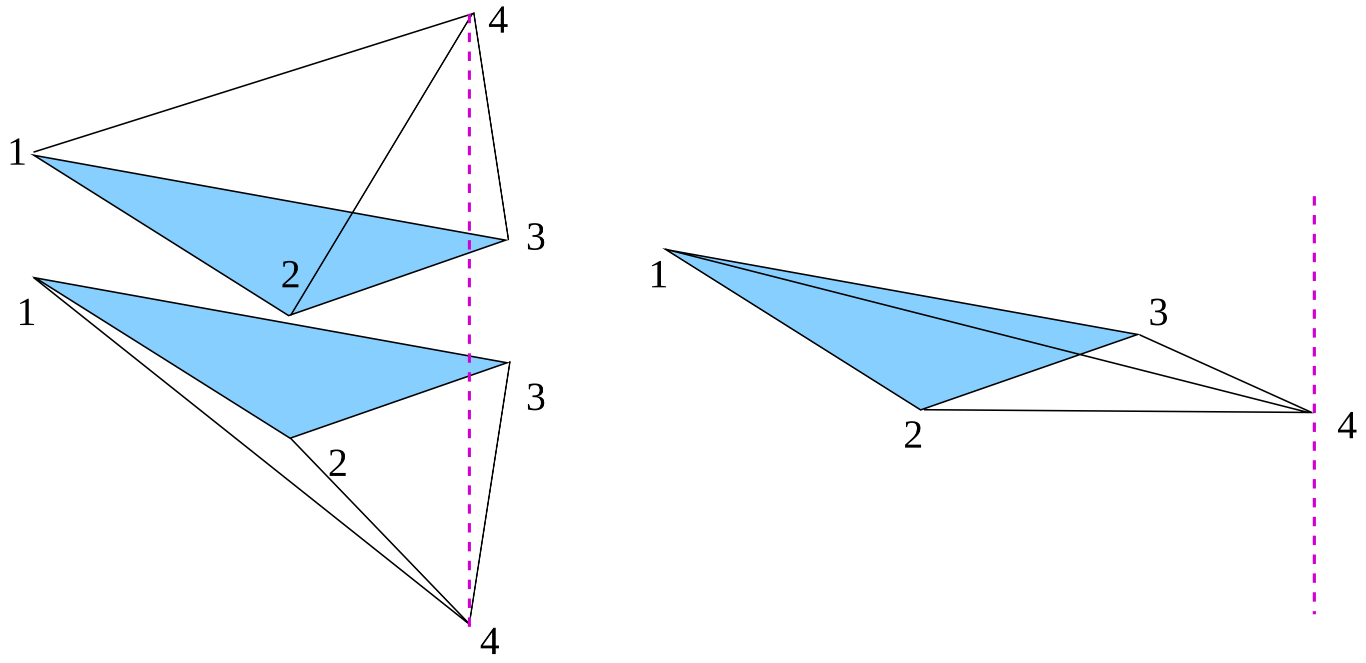
- (Exactly one solution for y_K) $\Leftrightarrow$ $\mu^2 = \lambda\nu$, not a tautology

- The set of all $(K+1)$-clique DGP instances in $\mathbb{R}^K$ s.t. $\mu^2 = \lambda\nu$ has Lebesgue measure 0

- **Ignore them, they happen with probability* 0!**

* Assuming continuous distributions over the reals. For floating point number, who knows? . . . but we'll ignore these instances anyhow

Discriminant > 0, $= 0$

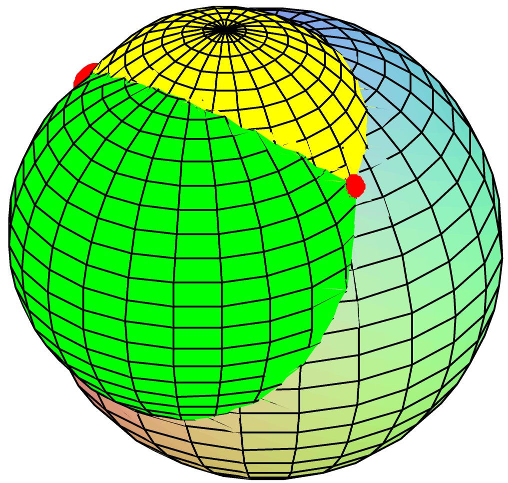


The case of two solutions

- K spheres $\mathbb{S}_1^{K-1}, \ldots, \mathbb{S}_K^{K-1}$ in $\mathbb{R}^K$

  centered at $x_1, \ldots, x_K$

  with radii $d_{1,K+1}, \ldots, d_{K,K+1}$
- x_{K+1} must be at the intersection of $\mathbb{S}_1^{K-1}, \ldots, \mathbb{S}_K^{K-1}$
- If $\bigcap_j \mathbb{S}_j^{K-1} \neq \varnothing$, then $|\bigcap_j \mathbb{S}_j^{K-1}| = 2$ in general



- *will not mention "probability 0" or "in general" anymore*

[Coope 2000]

Mirror images

- Let $x^+ = \{x_1, \ldots, x_K, x^+_{K+1}\}$, $x^- = \{x_1, \ldots, x_K, x^-_{K+1}\}$
  assume $\dim \operatorname{aff}(x_1, \ldots, x_K) = K$ (†)

- **Theorem**
  $x^+, x^- \in \mathbb{R}^K$ are reflections w.r.t. hyperplane defined by $x_1, \ldots, x_K$

- *Proof*
  1. x^+, x^- congruent by construction
  2. $\forall i \leq K \; x_i \in x^+ \cap x^- \rightarrow x^+, x^-$ not translations
  3. $|x^+ \cap x^-| = K < |x^+| = |x^-| \rightarrow x^+, x^-$ not rotations by (†)
  4. $\Rightarrow$ must be reflections

Algorithm for realizing $(K+1)$-cliques in $\mathbb{R}^K$

// realize 1 at the origin

$x_1 = (0, \ldots, 0)$

// realize next vertex iteratively

for $\ell \in \{2, \ldots, K+1\}$ **do**

    // at most two positions in $\mathbb{R}^{\ell-1}$ for vertex ℓ

    $S = \bigcap_{i<\ell} \mathbb{S}_i^{\ell-2}$

    if $S = \varnothing$ **then**

        // warn if infeasible

        return $\emptyset$

    end if

    // arbitrarily choose one of the two points

    choose any $x_\ell \in S$

end for

// return feasible realization

return x

Complexity of Alg. 2

- Outer loop: $O(K)$
- Gaussian elimination on A: $O(K^3)$
- Some messing about to obtain x^+_{K+1}, x^-_{K+1}: $+O(K^2)$
- Get $O(K^4)$
- But in most applications K is fixed
- **Get** $O(1)$

Back to complete graphs

- Alg. 2: realize $1, \ldots, K+1$ in $\mathbb{R}^K$: $O(1)$
- Alg. 1: Realize $K+2, \ldots, n$: $O(n^2)$
- $\Rightarrow O(n^2)$
- **What about $|X|$?**
  - Alg. 1 is deterministic: one solution from $x_1, \ldots, x_{K+1}$
  - Alg. 2 is stochastic: pick one of two values K times

$$\Rightarrow |X| = 2^K$$

K-trilaterative graphs

- In Alg. 1 we only need each $v > K+1$ to have $K+1$ adjacent predecessors in order to find a unique solution for x_v

- Determination of x_v from $K+1$ adjacent predecessors: *K-trilateration*

- *K-trilaterative graph*:
  (i) has a vertex order ensuring this property
  (ii) the initial $K+1$ vertices induce a $(K+1)$-clique
  the order is called *K-trilateration order*

- Alg. 1 realizes all K-trilaterative graphs

The DGP restricted to K-trilaterative graphs in $\mathbb{R}^K$ is easy

[Eren et al. 2004]

The story so far

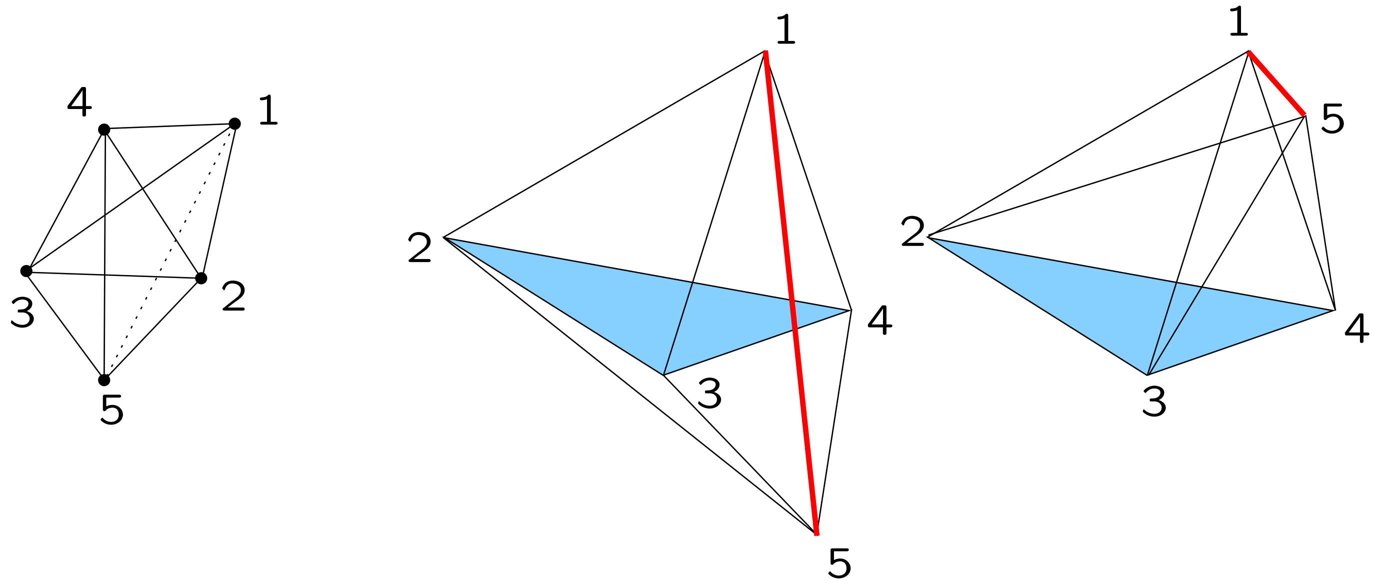
- Lots of nice applications
- DGP is **NP**-hard
- May have $0, 1$, finitely many or $2^{\aleph_0}$ solutions modulo congruences
- Continuous optimization techniques don't scale well
- Using $K + 1$ adjacent predecessors, realize K-trilaterative graphs in $\mathbb{R}^K$ in polytime
- **Do we** *need* $K + 1$ **adjacent predecessors, or can we do with less?**

The Branch-and-Prune algorithm

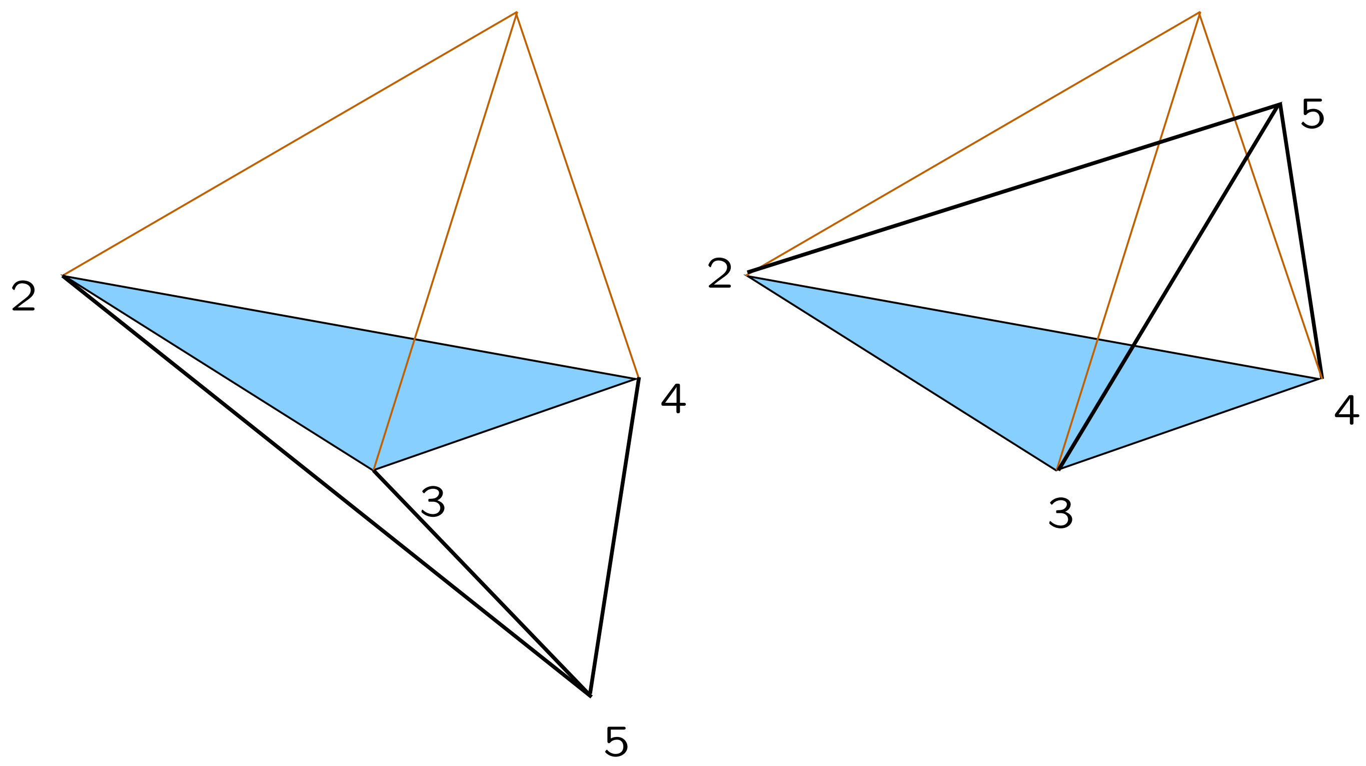
1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. **The Branch-and-Prune algorithm**
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

Fewer adjacent predecessors

- **Alg. 2 only needs K adjacent predecessor**
- Extend to n vertices: $(K-1)$-trilaterative graphs
- Can we realize $(K-1)$-trilaterative graphs in $\mathbb{R}^K$?
- *A small case: graph consisting of two $K+1$ cliques*



Take a closer look...



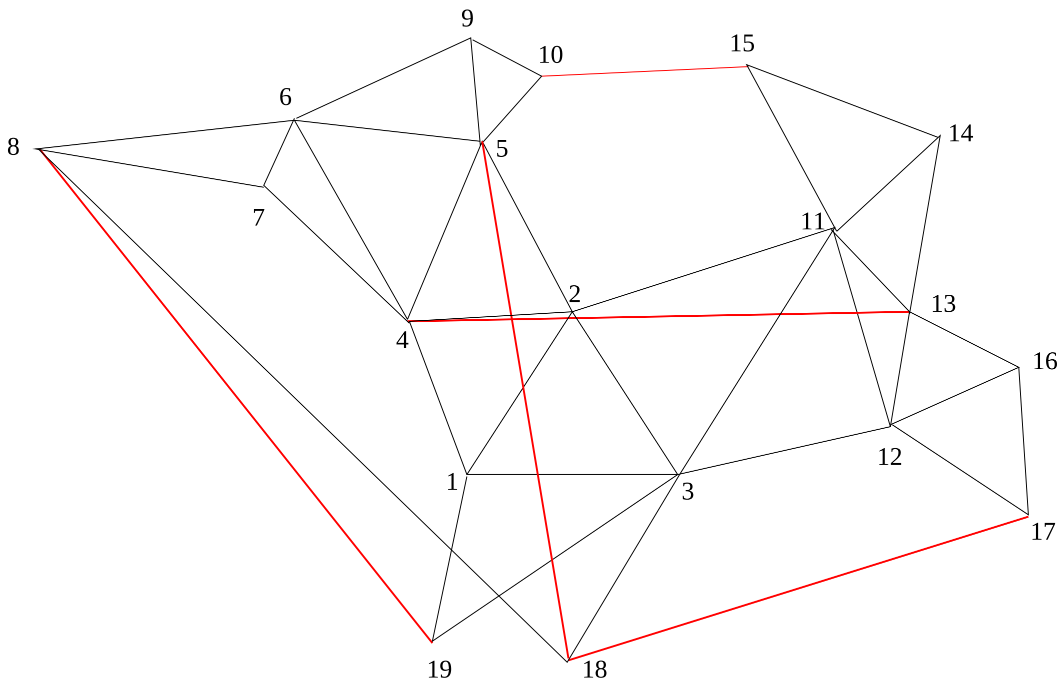
- Realization of a $K+1$ clique in $\mathbb{R}^K$ knowing $x_1, \ldots, x_K$
- **We know how to do that!**
- Consistent with 2 solutions for x_5, reflected across plane through x_2, x_3, x_4

Discretization and pruning edges

- $(K-1)$-*trilaterative graph* $G=(V,E)$:
  $\forall v > K \; \exists U_v \subset V \; (|U_v| = K \; \wedge \; \forall u \in U_v (u < v) \; \wedge \; \{u,v\} \in E)$

- **Discretization edges**:
  $E_D = \underbrace{\{\{u,v\} \in E \mid u,v \leq K\}}_{\text{initial clique}} \cup \underbrace{\{\{u,v\} \in E \mid v > K \wedge u \in U_v\}}_{\text{vertex order}}$

- **Pruning edges** $E_P = E \smallsetminus E_D$

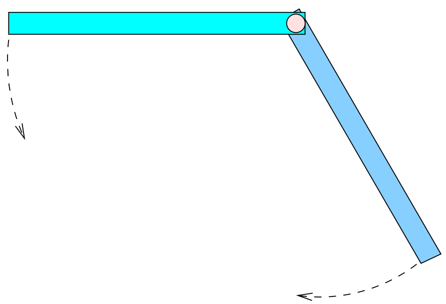


Role of discretization edges

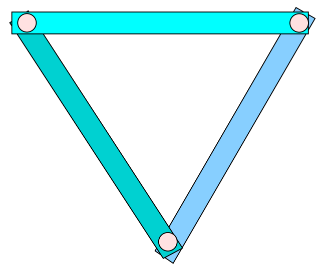
Missing discretization edge

$\Rightarrow$ non-rigid structure

$\Rightarrow$ X **uncountable**

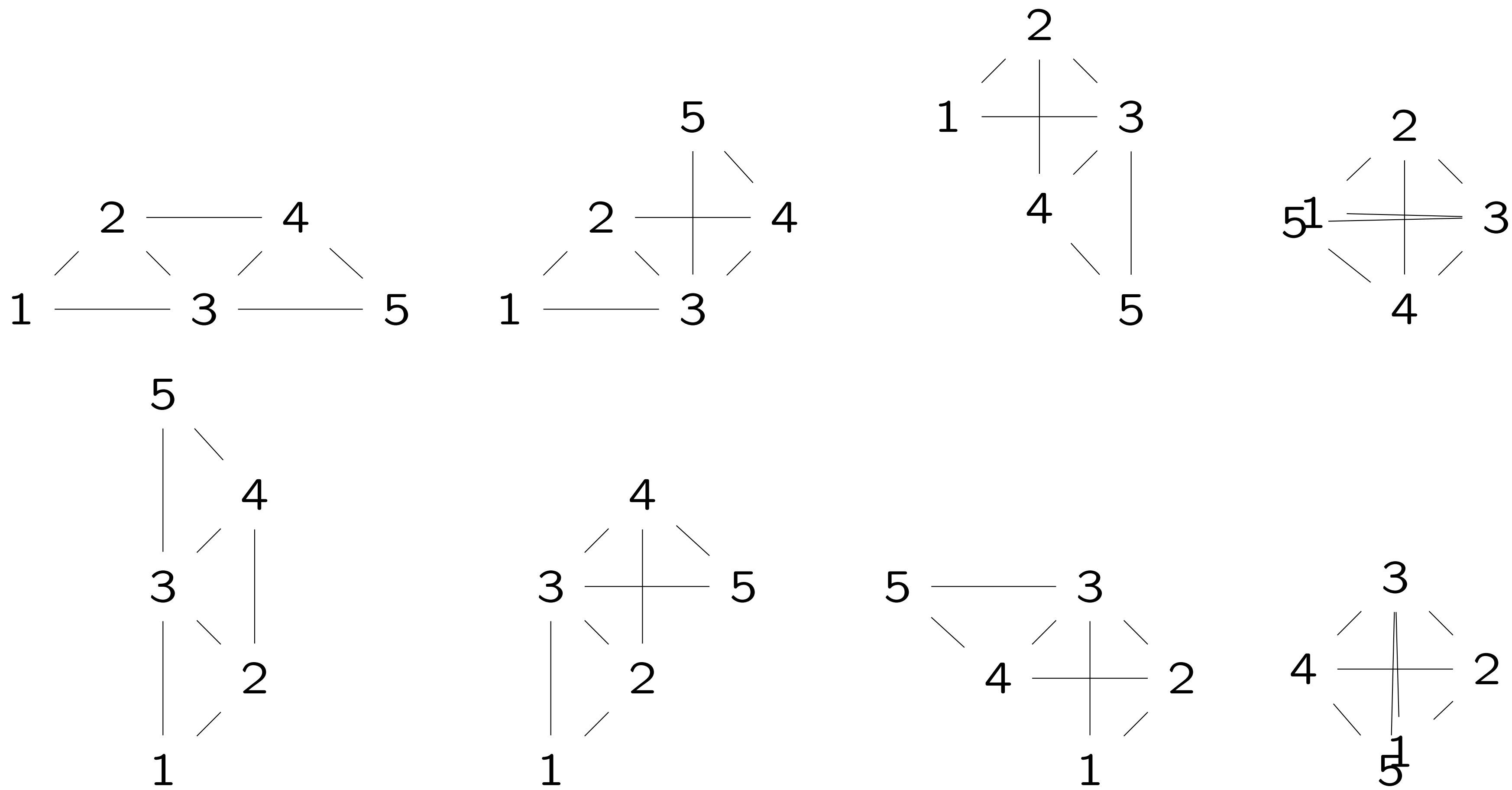


Else: X **finite**



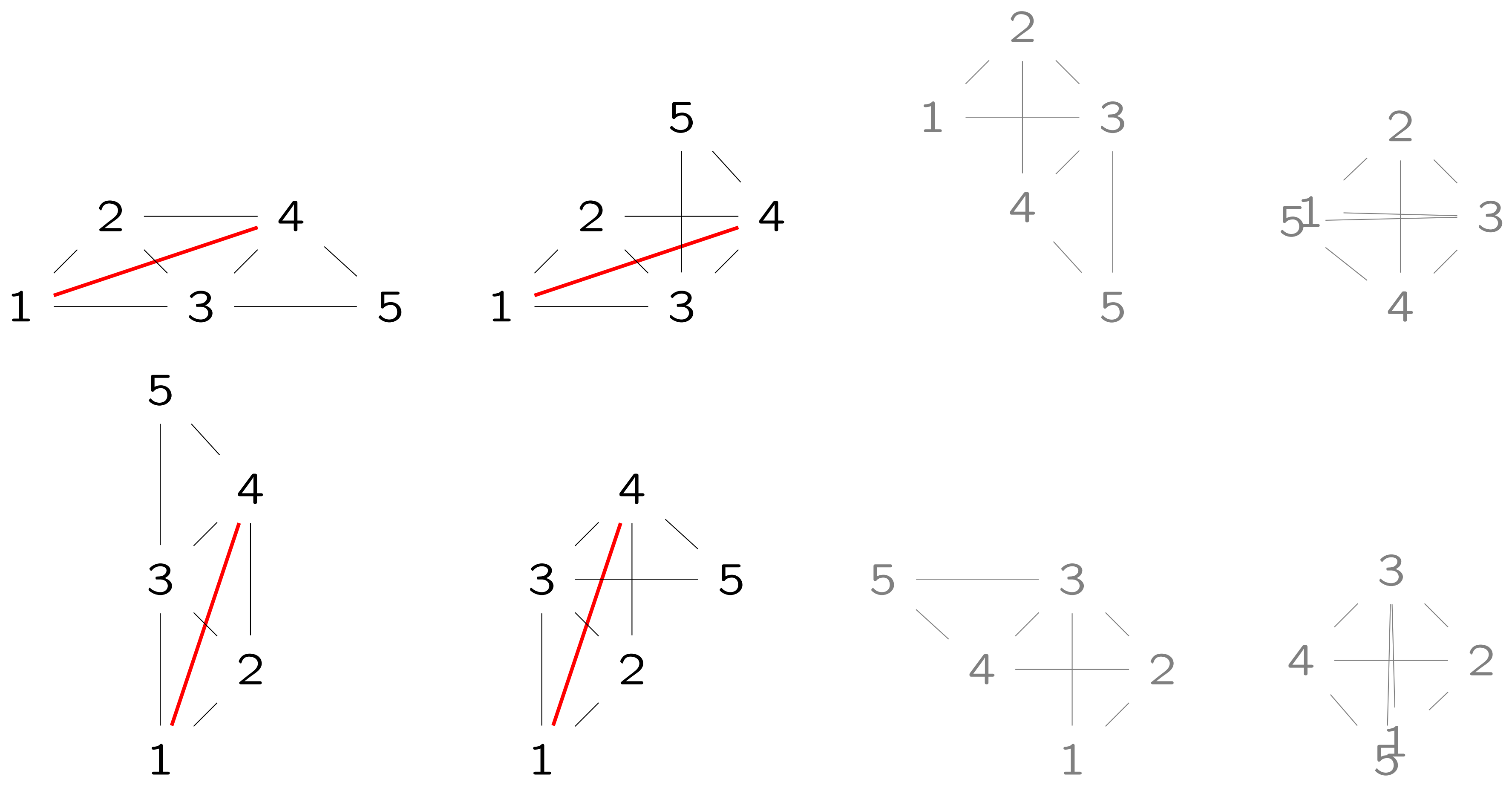
Role of pruning edges

No pruning edges: 8 incongruent realizations in $\mathbb{R}^2$



Role of pruning edges

Pruning edge $\{1, 4\}$: only 4 realizations remain valid



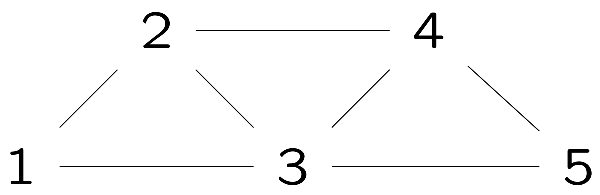
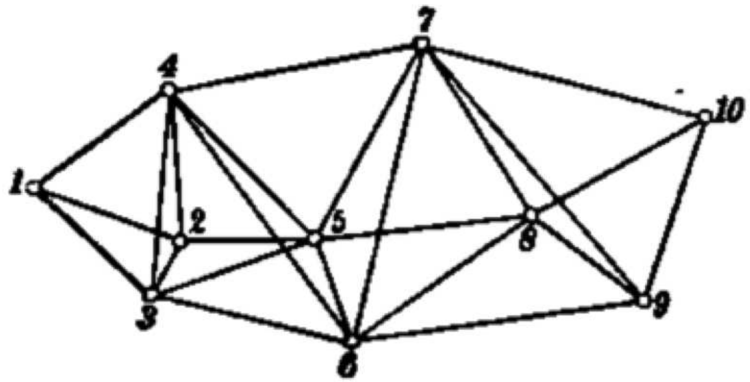
Protein backbones

- Total order $<$ on V
- Covalent bond **distances**: $\{u-1, u\} \in E$
- Covalent bond **angles**: $\{u-2, u\} \in E$
- **NMR experiments**: $\{u-3, u\} \in E$
  (and other edges $\{u, v\}$ with $v-u > 3$)

Generalize "3" to K

[Lavor et al., COAP 2012]

^{K}DMDGP graphs

$K = 2$	$K = 3$
	

Generalization of **protein backbone order**:

$v > K$ is adjacent to K **immediate predecessors** $v-1, \ldots, v-K$

^{K}DMDGP: Discretizable Molecular Distance Geometry Problem

The Branch-and-Prune (BP) algorithm

BP$(v, \bar{x}, X)$:

1. Given $v > K$, realization $\bar{x} = (x_1, \ldots, x_{v-1})$
2. Compute $S = \bigcap_{u \in U_v} \mathbb{S}_u^{K-1}$
3. For each $x_v \in S$ s.t. $\forall \{u, v\} \in E_P \ (u < v \to \|x_u - x_v\| = d_{uv})$
   (a) let $x = (\bar{x}, x_v)$
   (b) if $v = n$ add x to X, else call **BP**$(v + 1, x, X)$

- Recursive: starts with **BP**$(K + 1, (x_1, \ldots, x_K), \varnothing)$
- **All realizations in X are incongruent***
- Can be easily modified to find only p solutions for given p
- Applies to all $(K - 1)$-trilaterative graphs in $\mathbb{R}^K$
- Specialize to ^{K}DMDGP graph by setting $U_v = \{v - 1, \ldots, v - K\}$

* with probability 1, and aside from *one* reflection at $v = K + 1$

[L. et al. ITOR 2008]

Complexity of BP

- Most operations are $O(K^h)$ for some fixed $h \Rightarrow O(1)$
- Distance check at Step 3: $O(n)$
- Recursion on at most 2 branches at each call: **binary tree**
- Only recurse when $v > K, v < n$: 2^{n-K} nodes
- **Overall** $O(n2^{n-K}) = O(2^n)$

Worst-case exponential behaviour

Hardness of ^{K}DMDGP

- The ^{K}DMDGP is **NP**-hard for each K
  - every DGP instance is also DMDGP if $K = 1$
  - reduction from Partition can be extended to any K
- $(K-1)$-trilateration graphs are **NP**-hard by inclusion
- **No polytime algorithm unless P=NP**

Trilaterative graphs in $\mathbb{R}^K$ are complexitywise borderline at K

Correctness

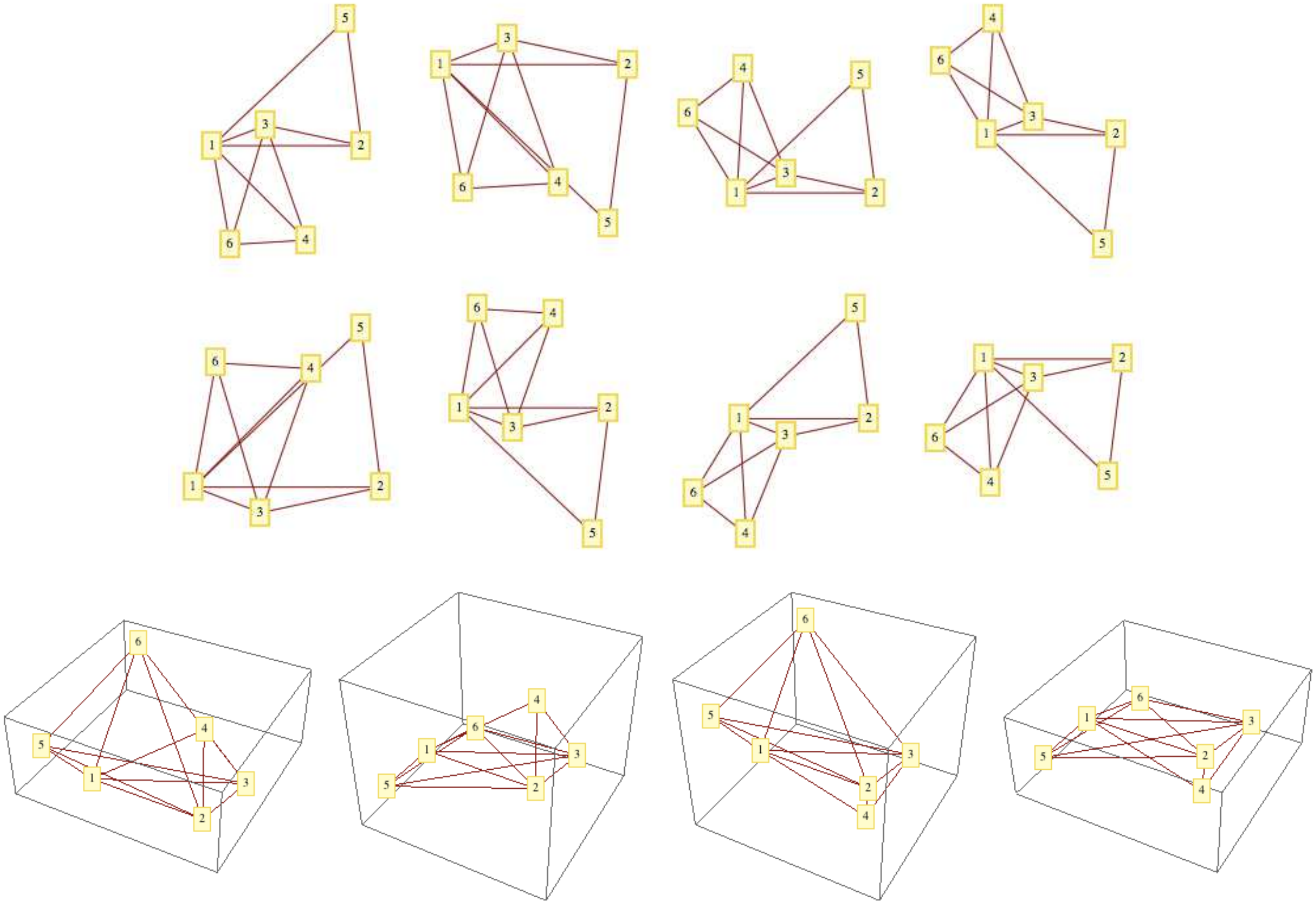
Thm.
When BP terminates, X contains every incongruent realization of G

Proof.

- Let $\bar{y}$ be any realization of G
- Since G has an initial K-clique, can rotate/translate/reflect $\bar{y}$ to $y[K] = x[K]$ for all $x \in X$
- BP exhaustively constructs every extension of $x[K]$ which is feasible with all distances, so $y \in X$

for a realization y, $y[h] = (y_1, \ldots, y_h)$ is the *initial segment* of y

Two examples



Empirical observations

- **Fast**: up to 10k vertices in a few seconds on 2010 hardware
- **Precise**: errors in range $O(10^{-9})$-$O(10^{-12})$
- Number of solutions always a power of 2:
  obvious if $E_P = \varnothing$, but otherwise mysterious
- **Linear-time behaviour on proteins**:
  this really shouldn't happen

Symmetry in the ^{K}DMDGP

1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. **Symmetry in the ^{K}DMDGP**
10. Tractability of protein instances
11. Finding vertex orders
12. Approximate realizations

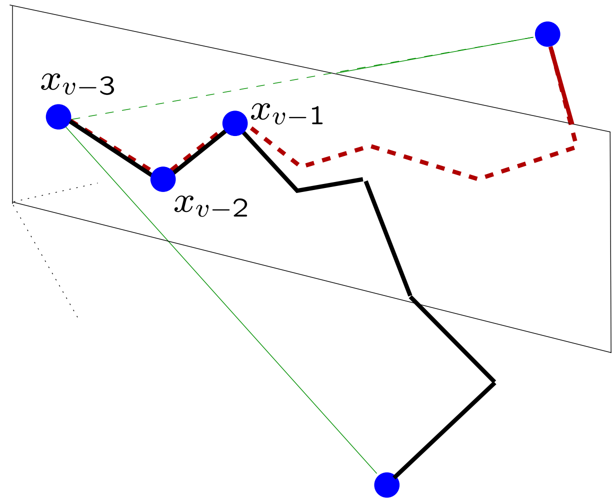
[L. et al. DAM 2014]

Partial reflections

- For each $v > K$, let

$$g_v(x) = (x_1, \ldots, x_{v-1}, R_x^v(x_v), \ldots, R_x^v(x_n))$$

be the *partial reflection* of x w.r.t. v



- Note: the g_v's are idempotent operators
- $G_D = (V, E_D)$: subgraph of G given by discretization edges
- $\forall v > K$ reflection R_x^v gives a binary choice in general*
- $X_D \subset \mathbb{R}^{nK}$ contains 2^{n-K} incongruent realizations of G_D

* subsequent results hold "with probability 1"

Discretization group

- $\mathscr{G}_D = \langle g_v \mid v > K \rangle$: the *discretization group* of G w.r.t. K
  subgroup of a Cartesian product of reflection groups

- An element $g \in \mathscr{G}_D$ has the form $\bigotimes_{v>K} g_v^{a_v}$, where $a_v \in \{0, 1\}$

- Action of $\mathscr{G}_D$ on X_D: $g(x) = \left(g_{K+1}^{a_{K+1}} \circ \cdots \circ g_n^{a_n}\right)(x)$

Commutativity of partial reflections

Lemma A $\mathscr{G}_D$ is Abelian

Proof Assume $K < u < v$. Then

$$\begin{aligned}
g_u g_v(x) &= g_u(x_1, \ldots, x_{v-1}, R^v_x(x_v), \ldots, R^v_x(x_n)) \\
&= (x_1 \ldots, x_{u-1}, R^u_{g_v(x)}(x_u), \ldots, R^u_{g_v(x)} R^v_x(x_v), \ldots, R^u_{g_v(x)} R^v_x(x_n)) \\
&= (x_1 \ldots, x_{u-1}, R^u_x(x_u), \ldots, R^v_{g_u(x)} R^u_x(x_v), \ldots, R^v_{g_u(x)} R^u_x(x_n)) \\
&= g_v(x_1, \ldots, x_{u-1}, R^u_x(x_u), \ldots, R^u_x(x_n)) \\
&= g_v g_u(x)
\end{aligned}$$

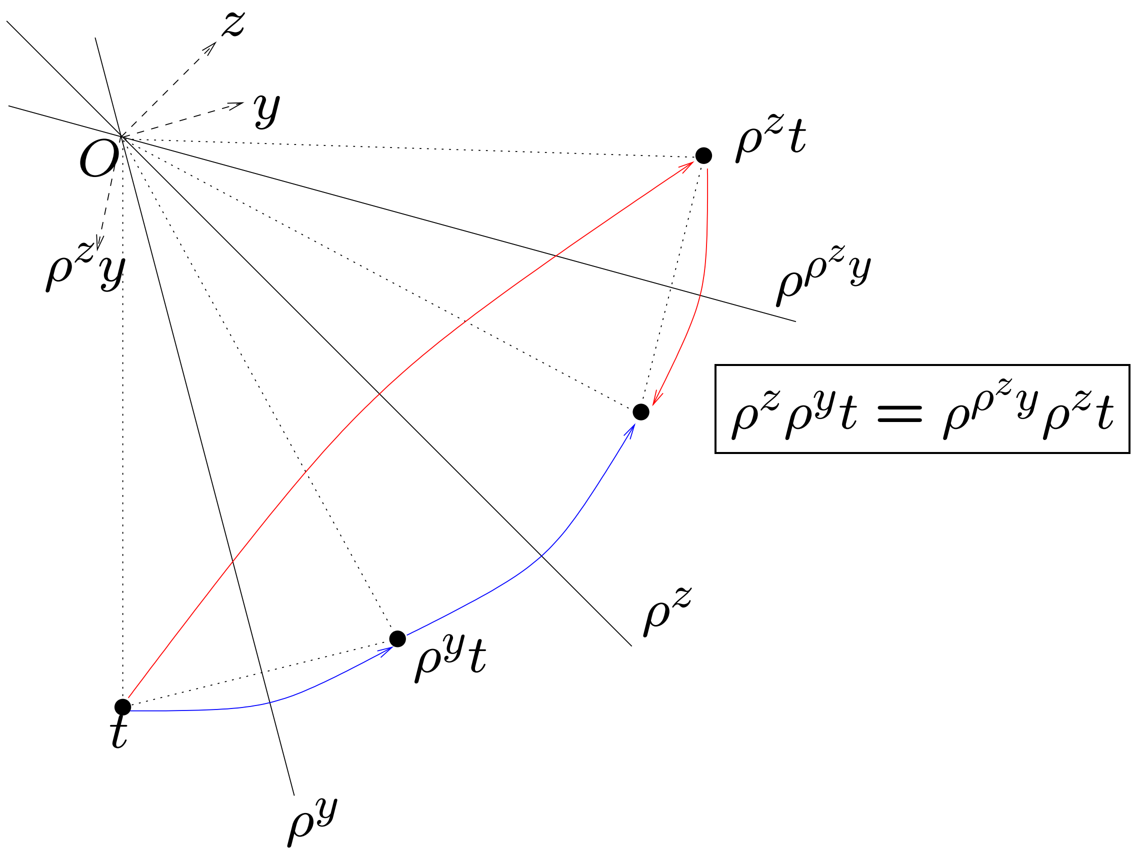
where equality of these terms holds by a Technical Lemma (next slide)

[L. et al. 2013]

Commutativity of partial reflections

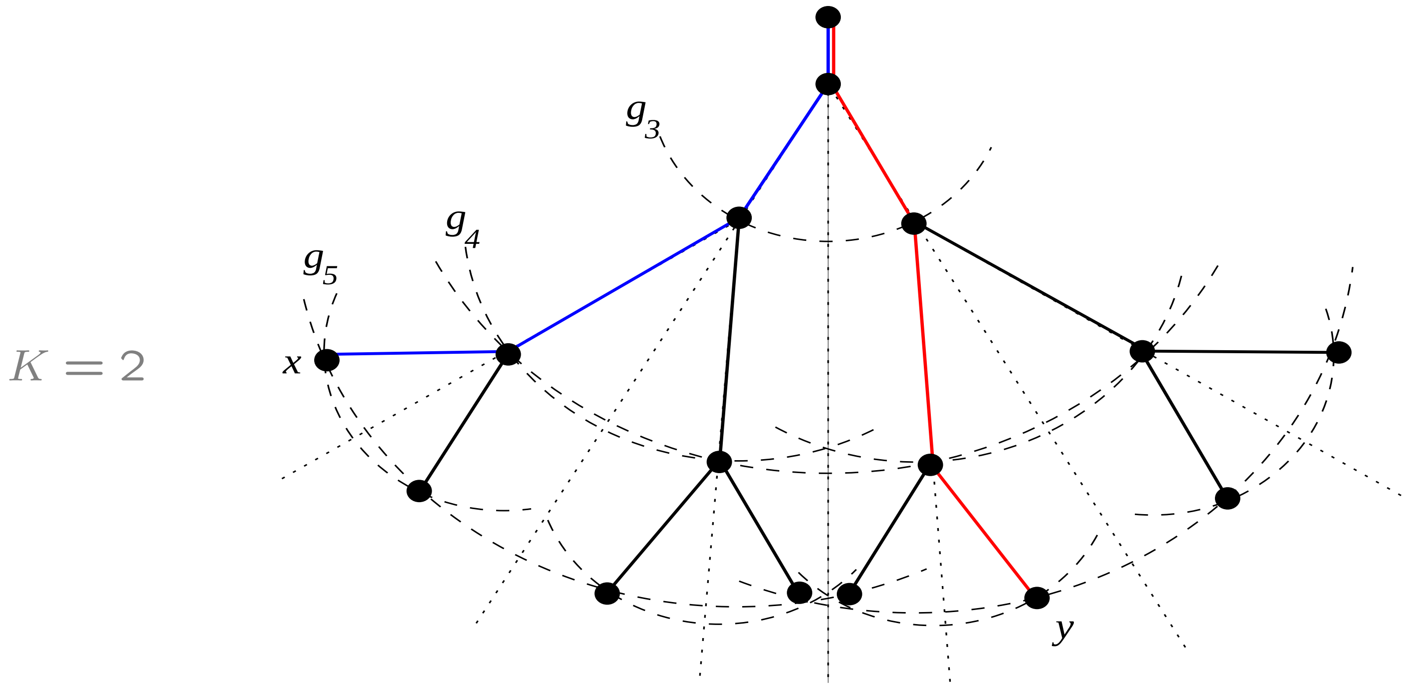
Technical Lemma

(Proof sketch for $K = 2$) Let $y \perp \mathrm{Aff}(x_{v-1}, \dots, x_{v-K})$ and $\rho^y = R_x^v$



One realization generates all others

Lemma B The action of $\mathscr{G}_D$ on X_D is transitive



$\exists g \in \mathscr{G}_D$ ($y = g(x)$): namely, $y = g_5(g_4(g_3(x)))$

Proof By induction on v: assume result holds to $v-1$ with g', then either it holds for v and $g = g'$, else flip and let $g = g_v g'$

[L. et al. 2013]

Structure and invariance

- $\mathscr{G}_D$ is Abelian and generated by $n-K$ idempotent elements

$$\Rightarrow \mathscr{G}_D \cong C_2^{n-K}$$

- $\mathscr{G}_D \leq \mathrm{Aut}(X_D)$ by construction

Solution sets

- X: set of incongruent realizations of G

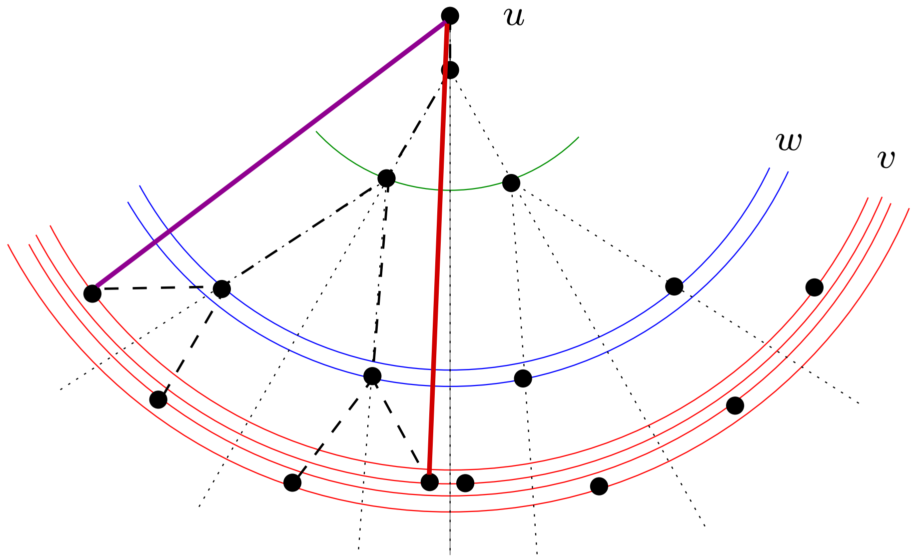
- G_D defined on same vertices but fewer edges
  $\Rightarrow$ fewer distance constraints on realizations
  $\Rightarrow$ more realizations

- All realizations of G are also realizations of G_D
  $\Rightarrow X \subseteq X_D$

Losing invariance on pruning edges

Lemma C Let $W^{uv} = \{u + K + 1, \ldots, v\}$ be the *range* of $\{u, v\}$
$\forall x \in X,\ u, w, v \in V\ (w \in W^{uv} \leftrightarrow \|x_u - x_v\| \neq \|g_w(x)_u - g_w(x)_v\|)$

Proof sketch for $K = 2$



Corollary If $\{u, v\} \in E_P$ and $w \in W^{uv}$, $g_w(x) \notin X$

[L. et al. 2013]

Pruning group

Define:

$$\begin{aligned} \Gamma &= \{g_w \in \mathscr{G}_D \mid w > K \wedge \forall \{u, v\} \in E_P \, (w \notin W^{uv})\} \\ \mathscr{G}_P &= \langle \Gamma \rangle \end{aligned}$$

Lemma D X is invariant w.r.t. $\mathscr{G}_P$

Proof

Follows by corollary, invariance of X_D w.r.t. $\mathscr{G}_D$ and $X \subseteq X_D$

Transitivity of the pruning group

Lemma E The action of $\mathscr{G}_P$ on X is transitive

- Given $x, y \in X$, aim to show $\exists g \in \mathscr{G}_P \; (y = g(x))$
- Lemma B $\Rightarrow \exists g \in \mathscr{G}_D$ with $y = g(x) \in X_D$
- Suppose $g \notin \mathscr{G}_P$ and aim for a contradiction
- $\Rightarrow \exists \{u, v\} \in E_P$ and $w \in W^{uv}$ s.t. g_w is a component of g
- Lemma C $\Rightarrow \|g_w(x)_u - g_w(x)_v\| \neq d_{uv}$
- If w is the only such vertex, $y = g(x) \neq x$ against hypothesis, done
- Suppose $\exists$ another $z \in W^{uv}$ s.t. g_z is a component of g
- Set of cases s.t. $\|x_u - x_v\| = \|g_z g_w(x)_u - g_z g_w(x)_v\|$ given $\|g_w(x)_u - g_w(x)_v\| \neq \|x_u - x_v\| \neq \|g_z(x)_u - g_z(x)_v\|$ has Lebesgue measure 0 in all DGP inputs
- By induction, holds for any number of components g_z of g with $z \in W^{uv}$
- $\Rightarrow y = g(x) \neq x$ against hypothesis, done

[L. et al. 2013]

The main result

Theorem $|X| = 2^{|\Gamma|}$

- Lemma A $\Rightarrow \mathscr{G}_D \cong C_2^{n-K} \Rightarrow |\mathscr{G}_D| = 2^{n-K}$
- $\mathscr{G}_P \leq \mathscr{G}_D \Rightarrow$ $\boxed{\exists \ell \in \mathbb{N}\ (\mathscr{G}_P \cong C_2^{\ell})}$, with $\ell = |\Gamma|$
- Lemma E $\Rightarrow \forall x \in X$ $\boxed{\mathscr{G}_P x = X}$
- Idempotency $\Rightarrow \forall g \in \mathscr{G}_P \quad g^{-1} = g$
  $\Rightarrow \forall g, h \in \mathscr{G}_P, x \in X\ (gx = hx \to h^{-1}gx = x \to hgx = x \to hg = I \to h = g^{-1} = g)$
  $\Rightarrow$ the mapping $\mathscr{G}_P x \to \mathscr{G}_P$ given by $gx \to g$ is injective
- $\forall g, h \in \mathscr{G}_P, x \in X\ (g \neq h \to gx \neq hx)$
  $\Rightarrow$ the mapping $gx \to g$ is surjective
- $\Rightarrow$ **the mapping $gx \to g$ is a bijection**
- $\Rightarrow |\mathscr{G}_P x| = |\mathscr{G}_P|$
- $\Rightarrow \forall x \in X \quad |X| = |\mathscr{G}_P x| = |\mathscr{G}_P| = 2^{|\Gamma|}$

[L. et al. 2013]

Symmetry-aware BP

- **Don't need to explore all branches of BP tree**
- Build Γ as a pre-processing step
- Run BP, terminating as soon as $|X| = 1$
- For each $g \in \mathscr{G}_P$, compute gx

[Mucherino et al. JBCB 2012]

Complexity

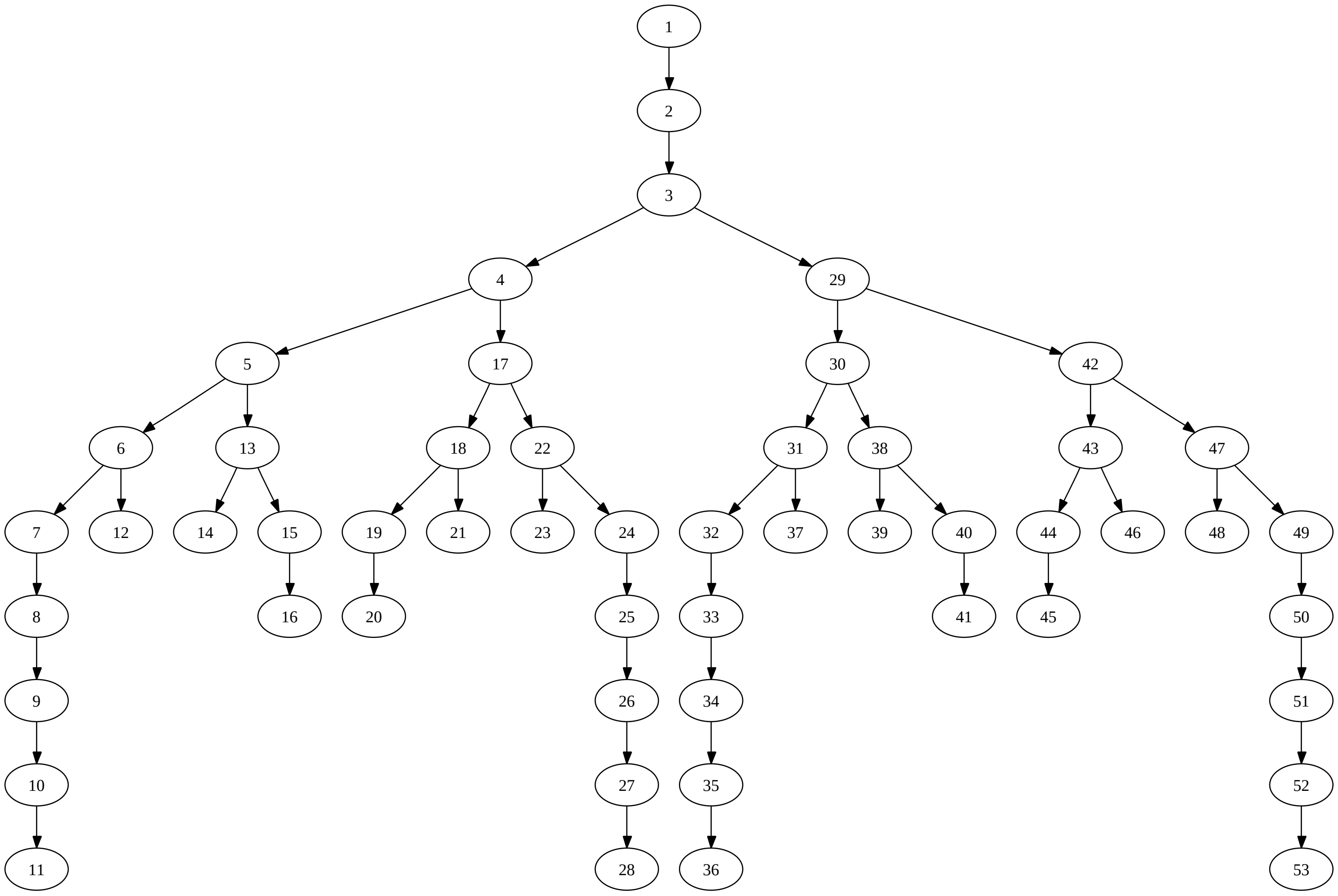
- Computing Γ: $O(mn)$
  1. initialize indicator vector $\iota = (\iota_{K+1}, \ldots, \iota_n)$ for $g_v \in \Gamma$
  2. initialize $\iota = \mathbf{1}$
  3. for each $\{u, v\} \in E_P$ and $w \in W^{uv}$ let $\iota_w = 0$
- BP: $O(2^n)$
- Compute gx for each $g \in \mathscr{G}_P$: $O(2^{|\Gamma|})$
- **Overall:** $O(2^n)$
- **Gains depend on the instance**

Tractability of protein instances

1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. **Tractability of protein instances**
11. Finding vertex orders
12. Approximate realizations

[L. et al. 2013]

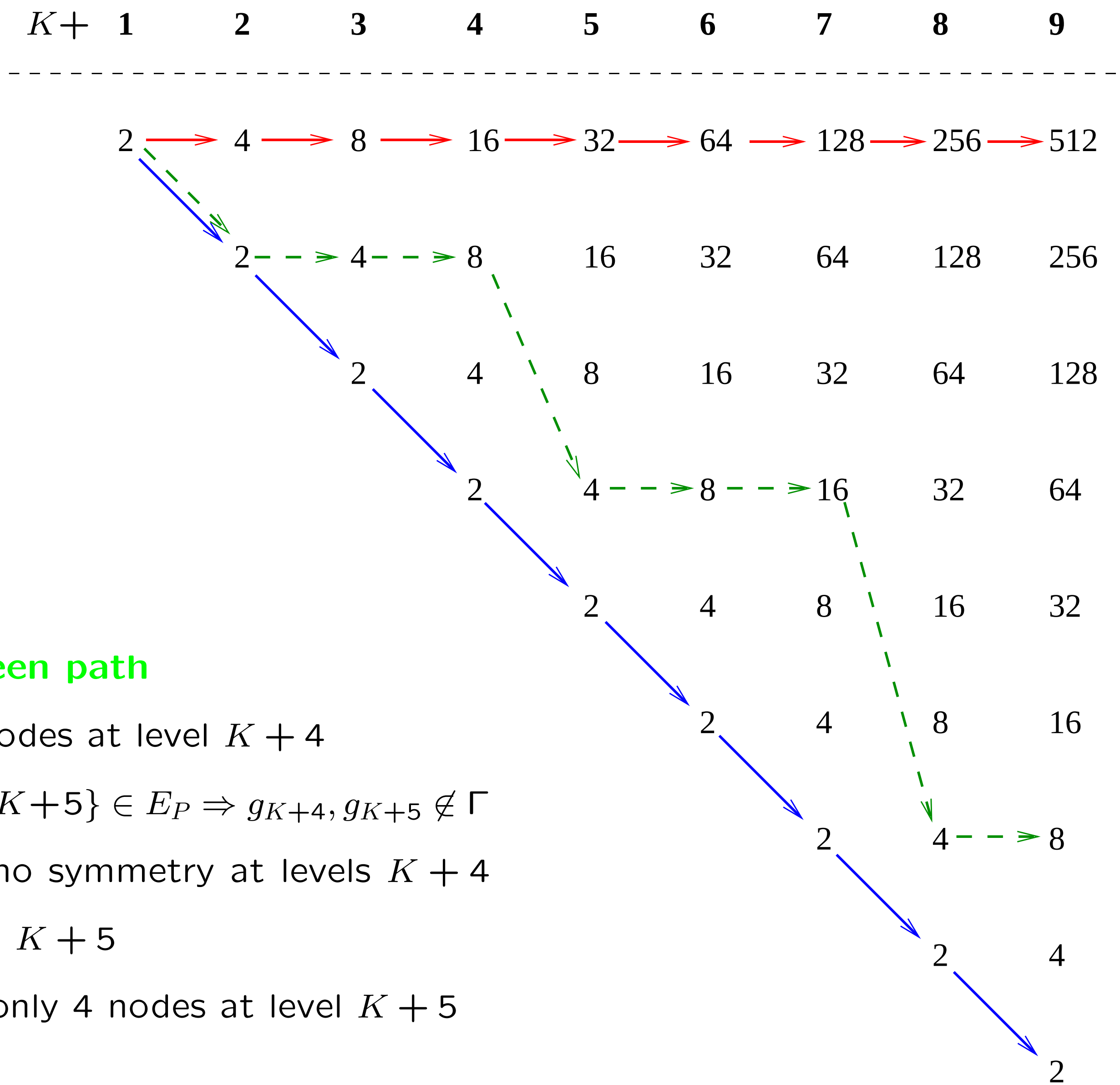
Let’s handle the BP tree



Max depth: n, looks good! Aim to prove width is bounded

Number of solutions at each BP tree level

Depends on range of longer pruning edge incident to level v



no pruning edges

Green path

8 nodes at level $K+4$

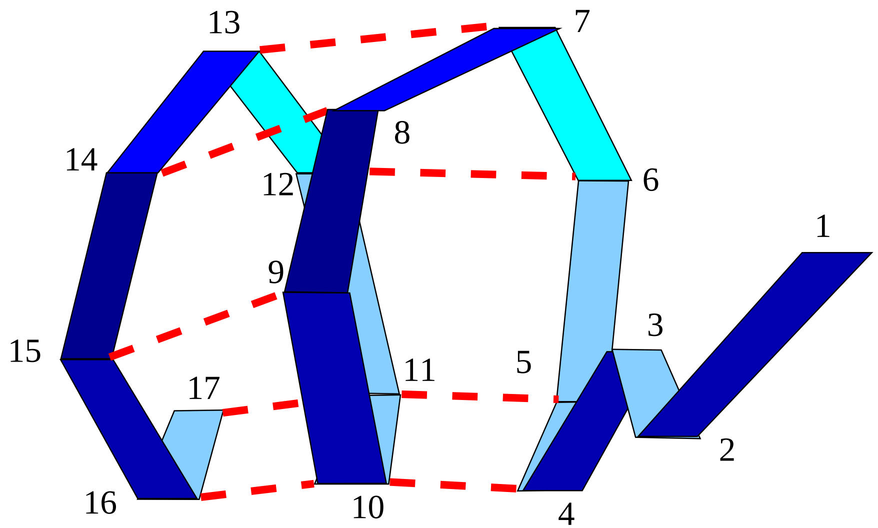
$\{3, K+5\} \in E_P \Rightarrow g_{K+4}, g_{K+5} \notin \Gamma$

$\Rightarrow$ no symmetry at levels $K+4$ and $K+5$

$\Rightarrow$ only 4 nodes at level $K+5$

pruning edges make graph K-trilaterative

Periodic pruning edges



4	5	6	7	8	9	10	11	12	...
2	4	8	16	32	64	128	256	512	
	2	4	8	16	32	64	128	256	
		2	4	8	16	32	64	128	
			2	4	8	16	32	64	
				2	4	8	16	32	
					2	4	8	16	
						2	4	8	
							2	4	
								2	

- 2^ℓ growth up to level ℓ, then constant: $O(2^\ell n)$ nodes in BP tree
- BP is **Fixed-Parameter Tractable** (FPT) in a bunch of cases
- For all tested protein backbones, $\ell \leq 5 \Rightarrow$ **BP linear on proteins!**

The story so far

- Nice applications, problem is hard, could have many solutions
- Continuous methods don't scale
- If *certain vertex orders* are present, use mixed-combinatorial methods
- Realize K-trilaterative in polytime but $(K-1)$-trilaterative are hard
- If adjacent predecessors are immediate, theory of symmetries
- Number of solutions is a power of two
- For proteins, BP is linear time
- **How do we find these vertex orders?**

Finding vertex orders

1. Applications
2. Definition
3. Complexity primer
4. Complexity of the DGP
5. Number of solutions
6. Mathematical optimization formulations
7. Realizing complete graphs
8. The Branch-and-Prune algorithm
9. Symmetry in the ^{K}DMDGP
10. Tractability of protein instances
11. **Finding vertex orders**
12. Approximate realizations

[Cassioli et al., DAM]

. . . wasn't the backbone providing them?

- NMR data not as clean as I pretended

- Have to mess around with side chains

- What about other applications, anyhow?

Methods for finding trilaterative orders automatically

Mostly bad news

- Finding K-trilaterative orders is **NP**-complete :-(
- **But also FPT :-)**
- Finding ^{K}DMDGP orders is **NP**-complete for all K :-(
- **It's also really hard in practice, and methods don't scale well**

Definitions

- Trilateration Ordering Problem (TOP)

  Given a connected graph $G = (V, E)$ and a positive integer K, does G have a K-trilateration order?

- Contiguous Trilateration Ordering Problem (CTOP)

  Given a connected graph $G = (V, E)$ and a positive integer K, does G have a $(K-1)$-trilateration order such that $U_v = \{v-1, \ldots, v-K\}$ for each $v > K$?

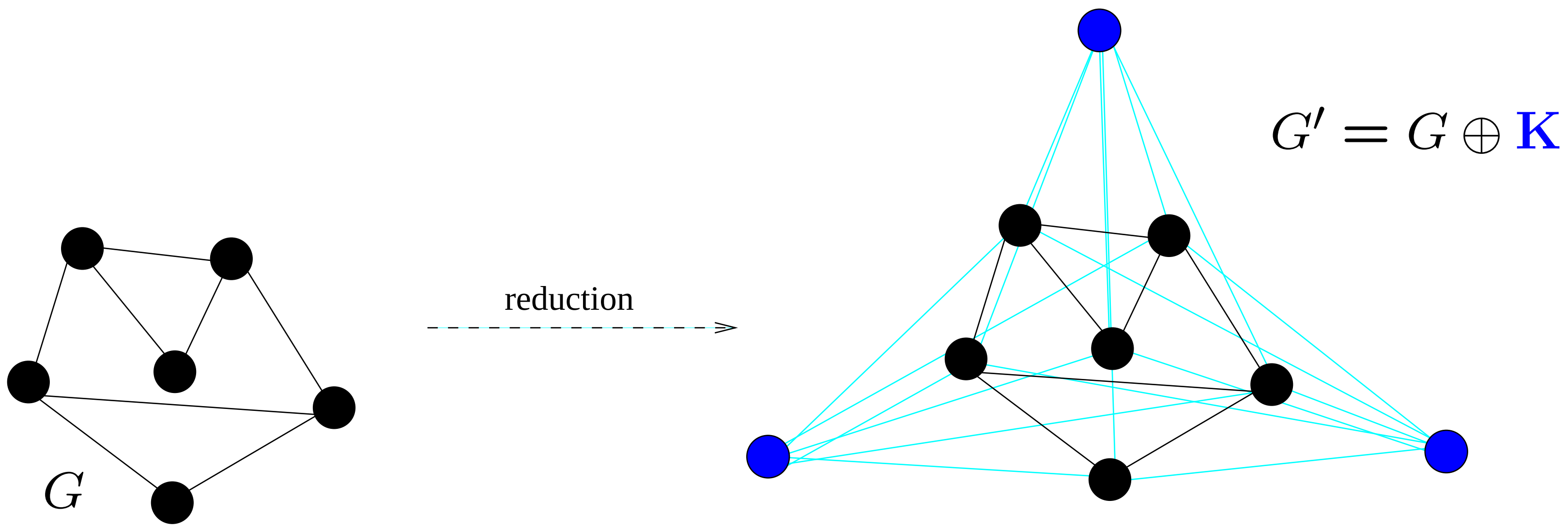
Both problems are in **NP**

Hardness of TOP

- Essentially due to finding the initial clique
  - brute force: test all $\binom{n}{K}$ subsets of V
  - $\binom{n}{K}$ is $O(n^K)$, polytime if K fixed
- Reduction from K-Clique problem:
  Given a graph, does it have a K-clique?

[Mucherino et al., OPTL 2012]

Reduction from K-Clique



- **If K-Clique instance is YES**
  - start with $\alpha =$ (initial clique of $G, \mathbf{K}$)
  - **induction**: if α_{v-1} defined, pick α_v at shortest path distance 1 from $\bigcup \alpha$
- **If K-Clique instance is NO**
  - **By contradiction**: suppose $\exists$ trilateration order α in G'
  - Initial clique $\alpha[K] = (\alpha_1, \ldots, \alpha_K)$ must have $K-1$ vertices in G, 1 in $\mathbf{K}$
  - α_{K+1} must be in G, hence $\exists$ K-clique in G

[Cassioli et al., DAM]

Once the initial clique is known

Greedily grow a trilateration order α

- Initialize α with initial K-clique $\mathbf{K}$
- Let $W = V \setminus \mathbf{K}$
- $\forall v > K \; a_v = |\text{vertices in } \mathbf{K} \text{ adjacent to } v|$
  // at termination, a_v will be the number of adjacent predecessors of v
- While $W \neq \varnothing$:
  1. choose $v \in W$ with largest a_v
  2. if $a_v < K$ instance is NO
  3. $\alpha \leftarrow (\alpha, v)$
  4. for all $u \in W$ adjacent to v, increase a_u
  5. $W \leftarrow W \setminus \{v\}$
- Instance is YES

[Mucherino et al., OPTL 2012]

Greedy algorithm is correct

- **Assume TOP instance is YES, proceed by induction**
  - start: by maximality, $a_{K+1} > K$
  - assume α is a valid TOP up to $v-1$, suppose $a_v < K$
  - but instance is YES so there is another $z \in W$ with $a_z \geq K$
  - contradicts maximality of a_v
- **Assume TOP instance is NO**
  - algorithm termination at $W = \varnothing$ contradicts the NO
  - hence it must terminate with $W \neq \varnothing$ and "NO" answer

Complexity

- Outer *while* loop: $O(n)$
- Choice of largest a_v: $O(n)$
- Inner loop on W: $O(n)$
- **Overall:** $O(n^2)$
- **If we add brute force initial clique:** $O(2^K n^2)$
- Polytime if K fixed, FPT otherwise

CTOP is hard

- Reduction from Hamiltonian Path (HP)

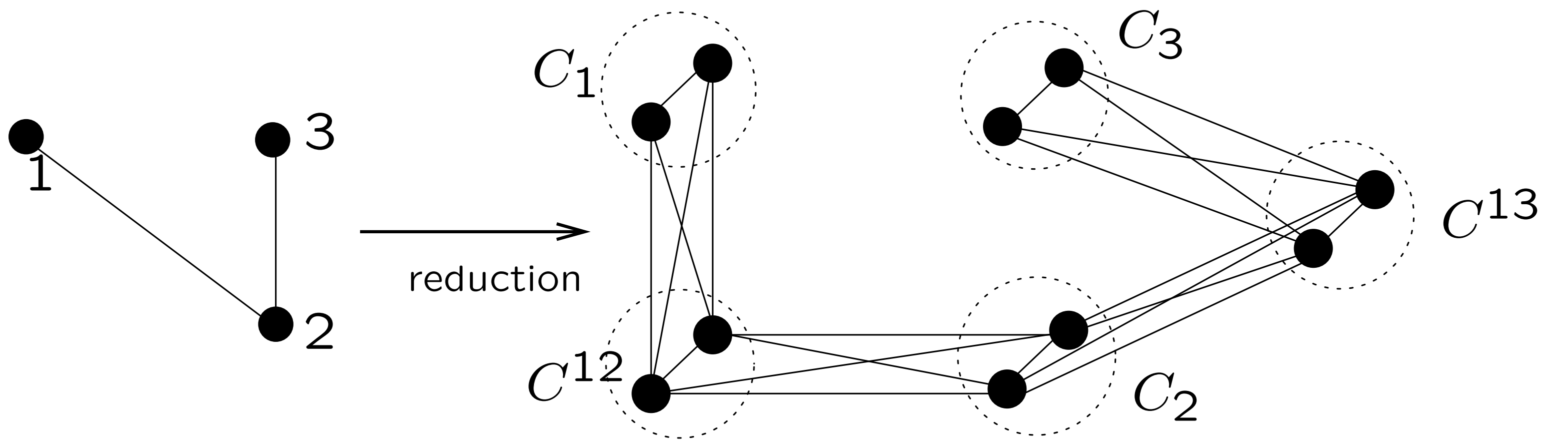
  Given a graph G, does it have a path passing through each vertex exactly once?

- α a H. path in $G \Rightarrow \forall v \neq 1, n$ α_v is adjacent to $\alpha_{v-1}, \alpha_{v+1}$
- Apart from initial 1-clique α_1
  every α_v is adjacent to its immediate predecessor
- $\Rightarrow$ α is a ^{K}DMDGP order in G with $K = 1$
- **HP is the same as ^{K}DMDGP with $K = 1$**
- $\Rightarrow$ **By inclusion, ^{K}DMDGP is NP-hard**

[Cassioli et al., DAM]

CTOP is hard for all K

- Reduction from HP



- Technical proof

[Cassioli et al., DAM]

How do we find ^{K}DMDGP orders?

Mathematical optimization & CPLEX

- $x_{vi} = 1$ iff vertex v has rank i in the order
- Each vertex has a unique order rank:

$$\forall v \in V \qquad \sum_{i \in \bar{n}} x_{vi} = 1;$$

- Each rank value is assigned a unique vertex:

$$\forall i \in \bar{n} \qquad \sum_{v \in V} x_{vi} = 1;$$

- There must be an initial K-clique:

$$\forall v \in V, i \in \{2, \ldots, K\} \qquad \sum_{u \in N(v)} \sum_{j < i} x_{uj} \ge (i-1) x_{vi};$$

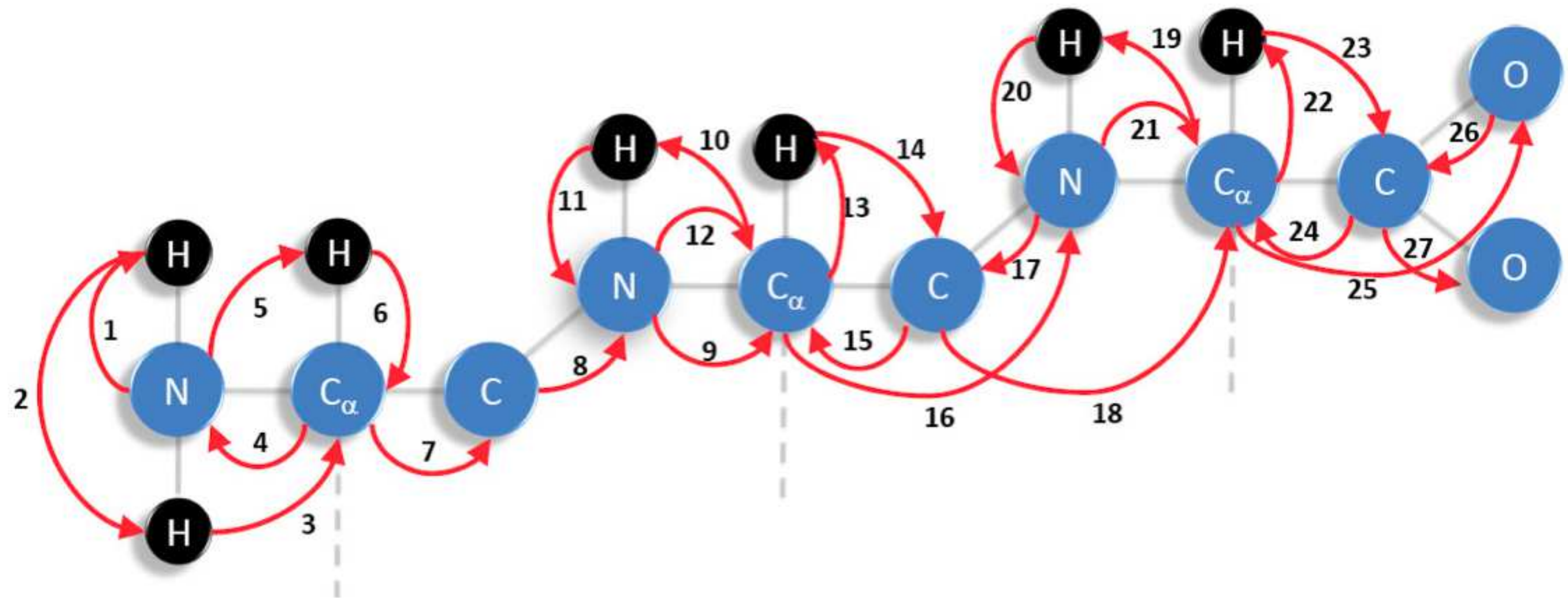
- Each vertex with rank $> K$ must have at least K contiguous adjacent predecessors

$$\forall v \in V, i > K \qquad \sum_{u \in N(v)} \sum_{i-K \le j < i} x_{uj} \ge K x_{vi}.$$

- Do not expect too much; scales up to 100 vertices

How about those 10k-atom backbones?

We have Carlile for those

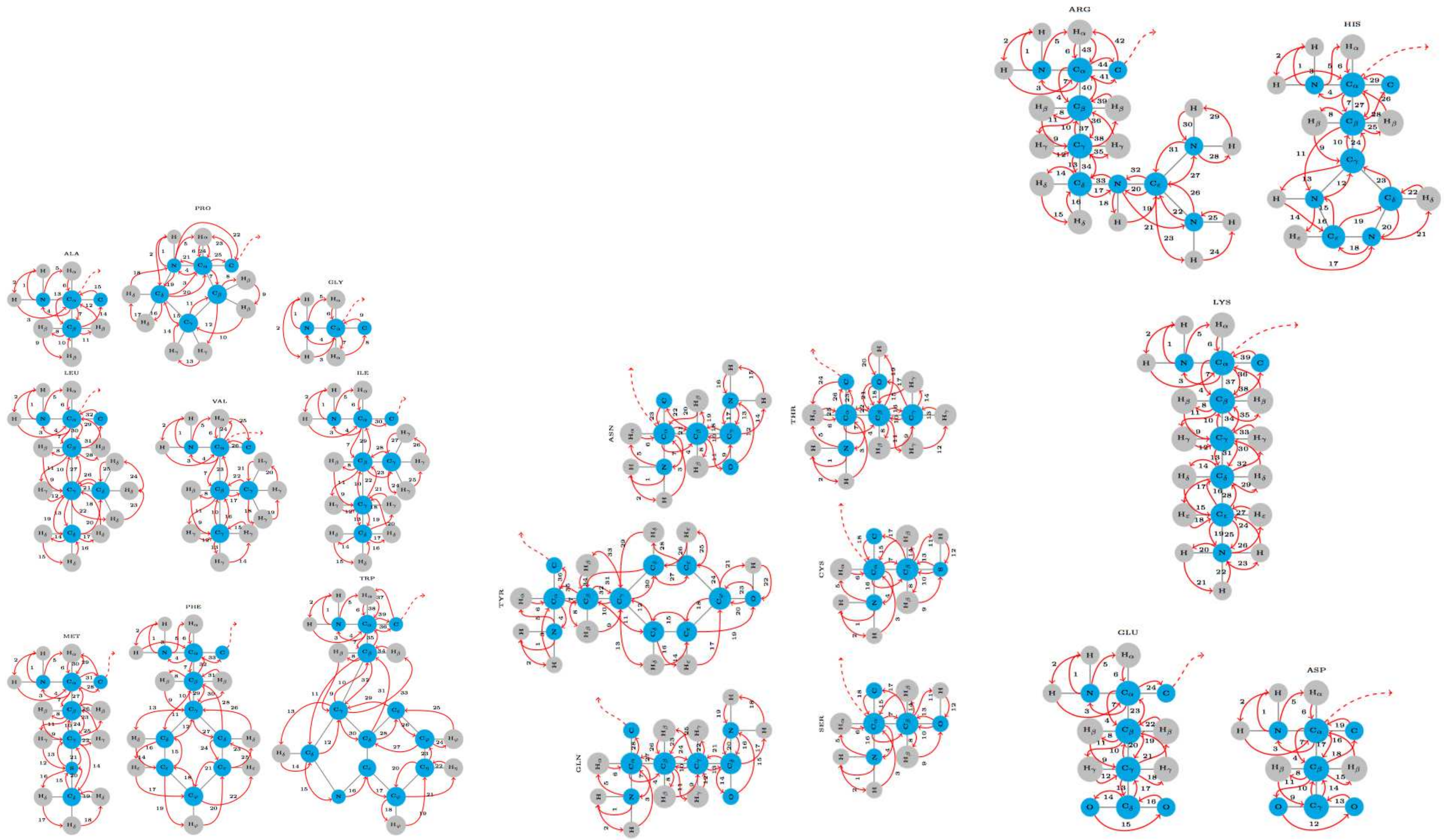


- Note the **repetitions** — they serve a purpose!
- Repetition orders are also hard to find for any K
- **. . . but Carlile knows how to handcraft them!**

[Lavor et al. JOGO 2013]

And what about the side-chains?

The Carlile+Antonio tool!



[Costa et al. JOGO, submitted]

Approximate realizations

Data errors

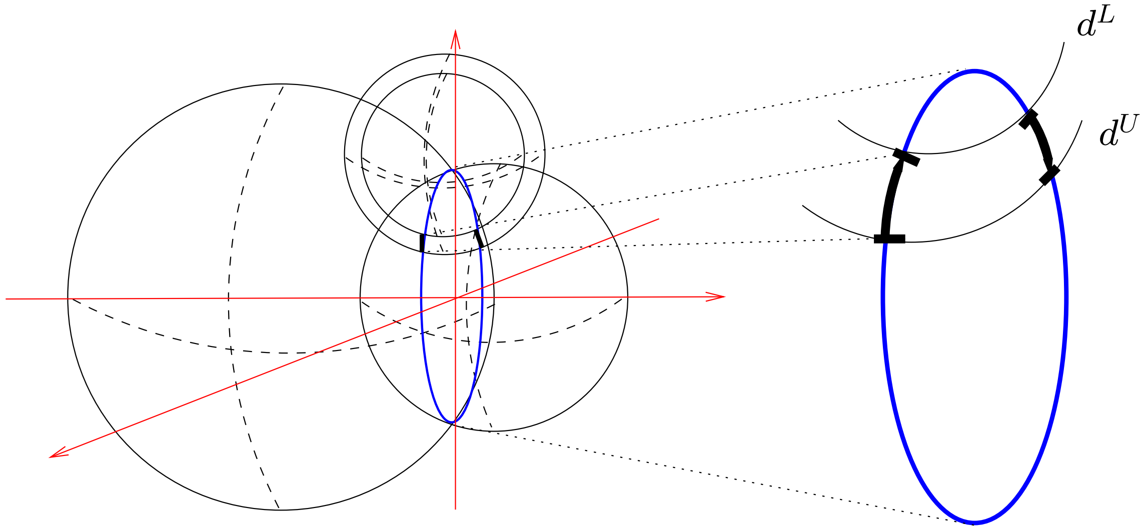
The "distance = real number" paradigm is a lie!

- Covalent bonds are fairly precise
- **NMR data is a mess** [Berger, J. ACM 1999]
  - experimental errors yield intervals $[d^L_{uv}, d^U_{uv}]$
  - NMR outputs frequencies of (atom type pair, distance value) weighted graph reconstruction yields systematic error
  - some atom type pairs yield more error ("only trust H—H")
- Properties of specific molecules give rise to other constraints
- **The protein graph may not be $(K-1)$-trilaterative based on the backbone**

The *Lavorder* comes to the rescue!

- **Carlile's handcrafted repetition orders properties**:
  - repetitions allow a "virtual backbone" of H atoms only
  - **discretization edges**: $\{v, v-i\}$ covalent bonds for $i \in \{1, 2\}$, $\{v, v-3\}$ sometimes covalent sometimes from NMR
  - most NMR data restricted to pruning edges

- When $d_{v,v-3}$ is an interval: intersect two spheres with sph. shell

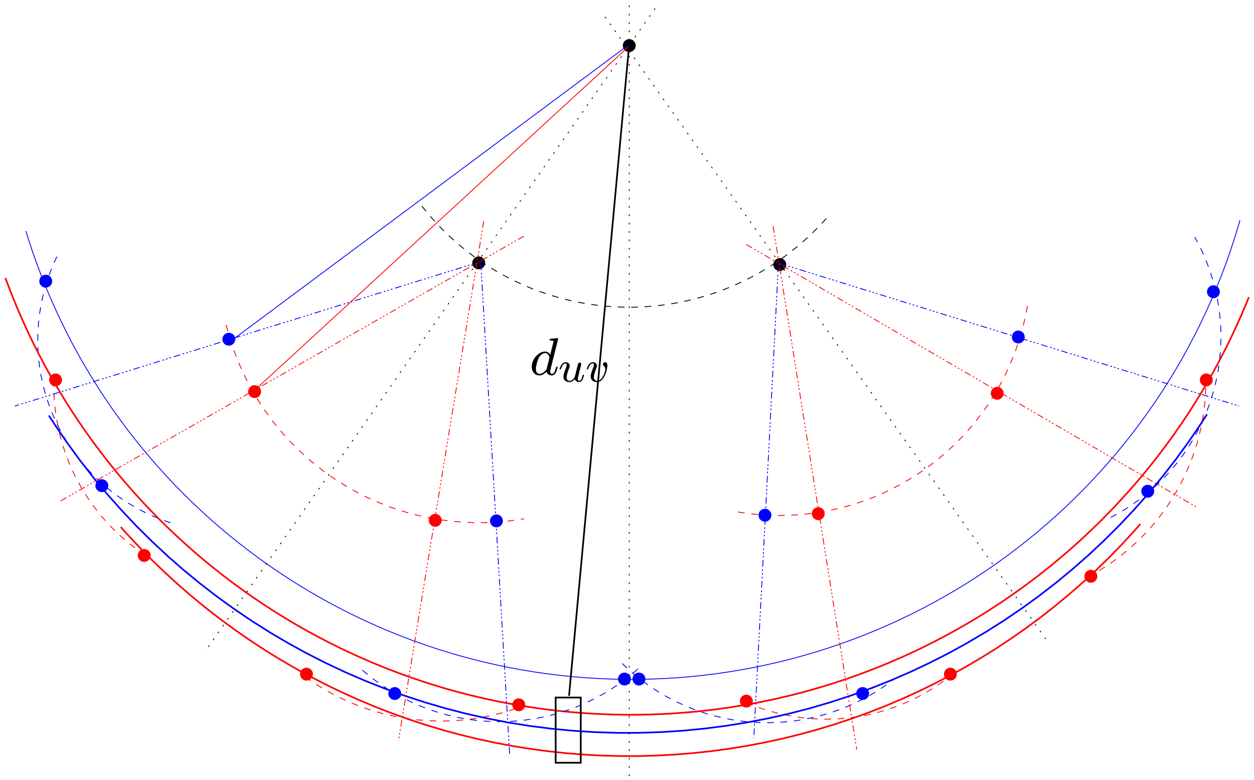


- Discretize circular segments and run BP with modified S

Algorithm no longer exhaustive

Die Symmetriktheoriedämmerung

- Intervals and discretization break the theory of symmetries

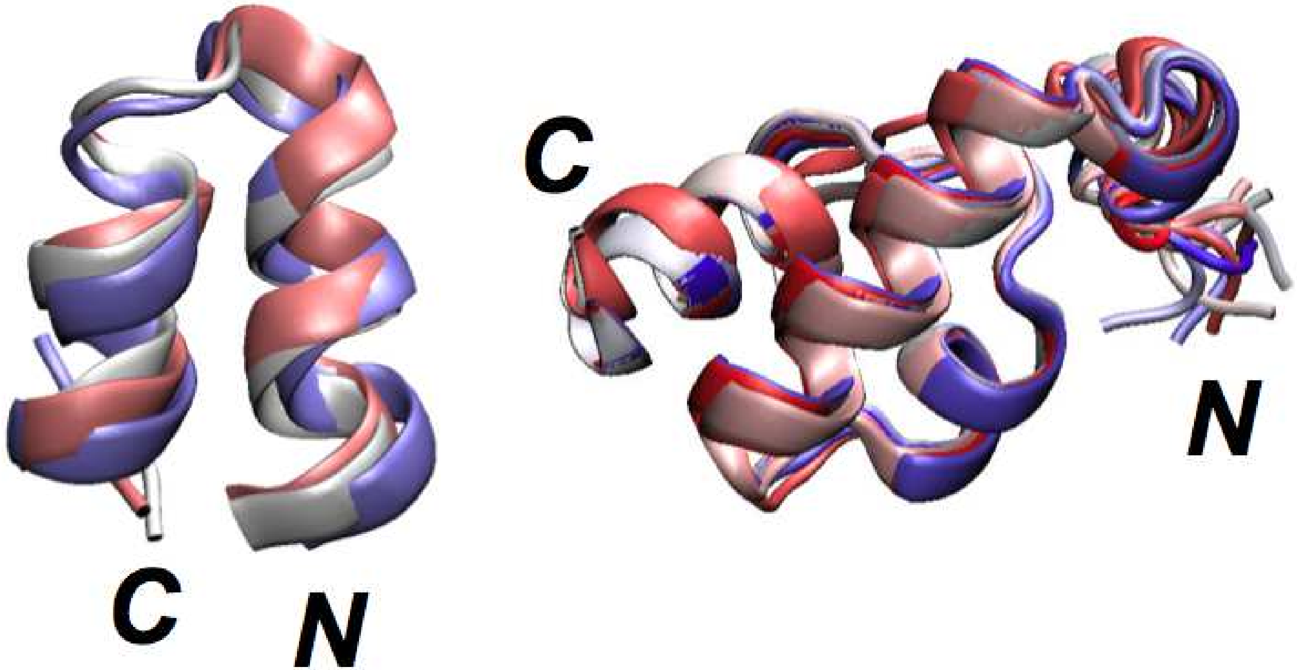


- Only some bounds for the number b of BP solutions:

$$\exists \ell, k \quad 2^{\ell} q^{k} \leq b \leq 2^{n-3} q^{M}$$

$q = |\text{discretization points}|$, $M = |\text{NMR discretization edges}|$

But at least it's producing results



Joint work with Institut Pasteur

[Cassioli et al., BMC Bioinf., submitted]

General approximate methods

- **All these methods are specialized to protein distance data from NMR**

- **What about general approximate methods?**

- Assume large-sized input data with errors

- No assumptions on graph structure

Ingredients

- PDM = Partial Distance Matrix (a representation of G)
- EDM = Euclidean Distance Matrix

1. **Complete** the given PDM d to a symmetric matrix D
2. **Find** a realization x (in some dimension $\bar{K}$)
   s.t. the EDM ($\|x_u - x_v\|$) is "close" to D
3. **Project** x from dimension $\bar{K}$ to dimension K,
   keeping pairwise distances approximately equal

Completing the distance matrix

- $\forall \{u, v\} \notin E$ let D_{uv} = length of the shortest path $u \to v$
- Use Floyd-Warshall's algorithm $O(n^3)$

// $n \times n$ array D_{ij} to store distances
$D = 0$
for $\{i, j\} \in E$ **do**
    $D_{ij} = d_{ij}$
end for
for $k \in V$ **do**
    for $j \in V$ **do**
        for $i \in V$ **do**
            if $D_{ik} + D_{kj} < D_{ij}$ **then**
                // D_{ij} fails to satisfy triangle inequality, update
                $D_{ij} = D_{ik} + D_{kj}$
            end if
        end for
    end for
end for

Finding a realization

- Let's give ourselves many dimensions, say $\bar{K} = n$
- Attempt to find $x : V \to \mathbb{R}^n$ with $(\|x_u - x_v\|_2) \approx (D_{uv})$
- **If we had the *Gram matrix* B of x, then**:
  1. find eigen(value/vector) matrices Λ, Y of B
  2. since B is PSD, $\Lambda \geq 0 \Rightarrow \sqrt{\Lambda}$ exists
  3. $\Rightarrow B = Y \Lambda Y^\top = (Y\sqrt{\Lambda})(Y\sqrt{\Lambda})^\top$
  4. $x = Y\sqrt{\Lambda}$ is such that $xx^\top = B$
- **Can we compute B from D?**

Schoenberg's theorem

- **Standard method for computing B from D^2**
- Also known as classic MultiDimensional Scaling (MDS)
- Apply many algebraic manipulations to

$$d_{uv}^2 = \|x_u - x_v\|^2 = {x_u}^\top x_u + {x_v}^\top x_v - 2{x_u}^\top x_v$$

  where the centroid $\sum_{k \le n} x_{uk} = 0$ for all $u \le n$

- Get $B = -\frac{1}{2}(I_n - \frac{1}{n}\mathbf{1}_n)D^2(I_n - \frac{1}{n}\mathbf{1}_n)$, i.e.

$$x_u \cdot x_v = \frac{1}{2n}\sum_{k \le n}(d_{uk}^2 + d_{kv}^2) - d_{uv}^2 - \frac{1}{2n^2}\sum_{\substack{h \le n \\ k \le n}} d_{hk}^2$$

- D "approximately" EDM $\Rightarrow$ B "approximately" Gram

[Schoenberg, Annals of Mathematics, 1935]

Project to $\mathbb{R}^K$ for a given K

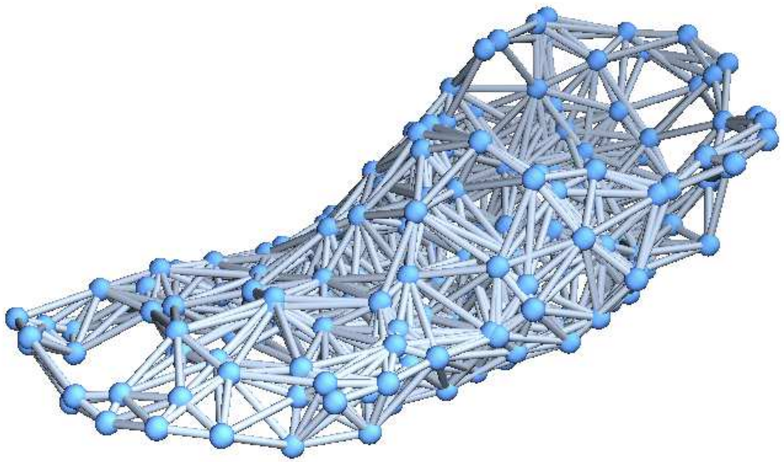
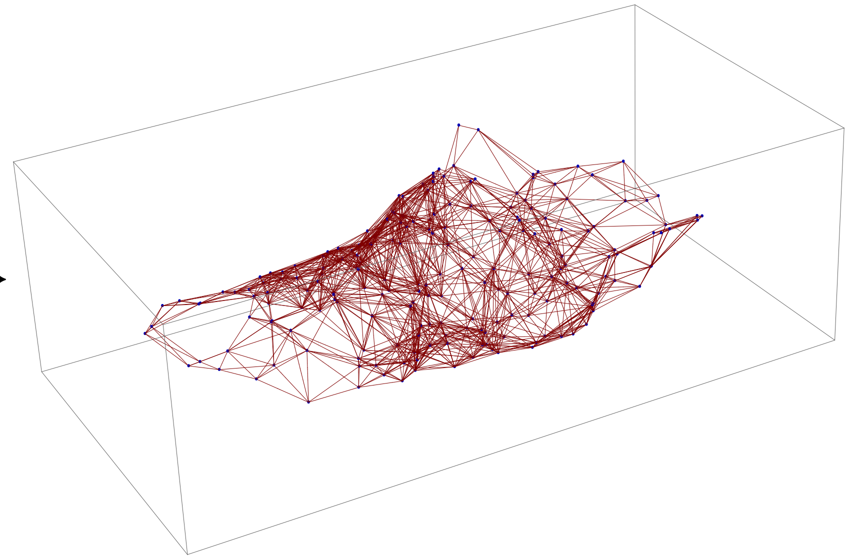
- Only use the K largest eigenvalues of Λ
- $Y[K] =$ K columns of Y corresp. to K largest eigenvalues
- $\Lambda[K] =$ K largest eigenvalues of Λ on diagonal
- $x = Y[K]\sqrt{\Lambda[K]}$ is a $K \times n$ matrix
- $Y[K]$ span the subspace where x “fills more space”, i.e. neglecting other dimensions causes smaller errors w.r.t. the realization in $\mathbb{R}^n$

This method is called **Principal Component Analysis** (PCA)

Isomap

Given K and PDM d:

1. $D = \texttt{FloydWarshall}(d)$
2. $B = \texttt{MDS}(D)$
3. $x = \texttt{PCA}(B, K)$

 $\longrightarrow$ 

[Tenenbaum et al. Science 2000]

Some references

- **L. Liberti**, C. Lavor, N. Maculan, A. Mucherino, *Euclidean distance geometry and applications*, SIAM Review, **56**(1):3-69, 2014

- **L. Liberti**, B. Masson, J. Lee, C. Lavor, A. Mucherino, *On the number of realizations of certain Henneberg graphs arising in protein conformation*, Discrete Applied Mathematics, **165**:213-232, 2014

- **L. Liberti**, C. Lavor, A. Mucherino, *The discretizable molecular distance geometry problem seems easier on proteins*, in [see below], 47-60

- A. Mucherino, C. Lavor, **L. Liberti**, N. Maculan (eds.), *Distance Geometry: Theory, Methods and Applications*, Springer, New York, 2013

THE END