Distance Geometry Problems

Leo Liberti

LIX, École Polytechnique, France

Joint work with:
C. Lavor (IMECC-UNICAMP), N. Maculan (COPPE-UFRJ), A. Mucherino (CERFACS)
At a glance

Which graph has most symmetries?
Which graph has most edges?
Is there a planar graph in this list?
Embed this graph in $\mathbb{R}^2$

- $G = (V, E, d)$ where $d : E \to \mathbb{Q}_+$

- $V = \{1, 2, 3, 4\}$

- $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{2, 4\}\}$

- $d(e) = 1$ for all $e \in E$

Infinitely many possible embeddings
What about this graph?

- $G = (V, E, d)$ where $d : E \to \mathbb{Q}_+$
- $V = \{1, 2, 3, 4\}$
- $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 4\}, \{2, 4\}\}$
- $d(\{1, 4\}) = 1.7$ and $d(\{e\}) = 1$ for all other $e \in E$

Two possible embeddings
Embedding protein data in $\mathbb{R}^3$

1aqr: four non-isometric embeddings
The basic problems
Problems in distance geometry

**Distance Geometry Problem (DGP)**

Given a simple, weighted undirected graph $G = (V, E, d)$ where $d : E \rightarrow \mathbb{Q}_+$ and $K \in \mathbb{N}$, is there an embedding $x : V \rightarrow \mathbb{R}^K$ such that $\forall \{u, v\} \in E \ (\|x_u - x_v\| = d_{uv})$?

(notation: $x_u = x(u)$ and $d_{uv} = d(\{u, v\})$)

**Euclidean Distance Matrix Completion Problem (EDMCP)**

Given a simple, weighted undirected graph $G = (V, E, d)$ where $d : E \rightarrow \mathbb{Q}_+$ and $K' \in \mathbb{N}$, is there an integer $K \leq K'$ and an embedding $x : V \rightarrow \mathbb{R}^K$ such that $\forall \{u, v\} \in E \ (\|x_u - x_v\| = d_{uv})$?

<table>
<thead>
<tr>
<th></th>
<th>DGP</th>
<th>EDMCP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Difference</strong></td>
<td>$K \in \text{input}$</td>
<td>$K \in \text{output}$</td>
</tr>
<tr>
<td><strong>Certificate</strong></td>
<td>$x$</td>
<td>$(K, x)$</td>
</tr>
</tbody>
</table>
Application to proteomics

Function of a protein: determined by its shape

Need to find positions of atoms in space

Some inter-atomic distances known from chemistry

Can compute some distances using NMR
  - shorter than 5Å
  - between pairs of H atoms
The molecule graph

- Set of atoms in the molecule: $V$
- Pairs of atoms for which we know a distance: $E$
- Distance function: $d : E \rightarrow \mathbb{Q}_+$
- Get weighted undirected graph $G = (V, E, d)$

Notation:

$$
\forall v \in V \quad \delta(v) = \{ u \in V \mid \{u, v\} \in E \} \quad : \text{adjacents (star)}
$$

$$
\forall U \subseteq V \quad E[U] = \{ \{u, v\} \in E \mid u, v \in U \} \quad : \text{edges within } U
$$

$$
\forall U \subseteq V \quad G[U] = (U, E[U]) \quad : \text{subgraph induced by } U
$$

Assume WLOG $G$ is connected
Example

An artificial protein test set: $\text{lavor-11}_7$
Other applications

- Graph Drawing
- Sensor Network Localization
Solutions modulo isometries

Suppose $x : V \rightarrow \mathbb{R}^K$ is a valid embedding

Let $T$ be an isometry

Then $T(x) : V \rightarrow \mathbb{R}^K$ is also a valid embedding

Thus, if an instance is YES, there are always infinitely many solutions

Compute “modulo isometries”

let $\bar{X} =$ set of valid embeddings

for all $x, y \in \bar{X}$ define $x \sim y$ if $\exists$ isometry $T$
s.t. $y = T(x)$

$\sim$ is an equivalence relation

We are only interested in $X = \bar{X}/\sim$
Main variants

- **DGP:**
  1. $\text{DGP}_K$: like the DGP, but $K$ is fixed and is not part of the input
  2. **Molecular Distance Geometry Problem (MDGP)** = DGP$_3$

- **EDMCP:**
  1. **Positive Semi-Definite Matrix Completion Problem (PSDMCP):** given a simple, weighted, undirected graph $G = (V, E, d)$ where $d : E \to \mathbb{R}_+$ is there a $|V| \times |V|$ positive semidefinite matrix $D = (D_{ij})$ such that $\forall \{i, j\} \in E \ (D_{ij} = d_{ij})$?
  2. Other variants with other types of matrices (see [Laurent 09])

A matrix $D$ is PSD if $\forall z \in \mathbb{R}^{|V|} \ (z^T D z \geq 0)$
**Variants: notes**

**DGP and DGP\(_K\): how do they differ?**

An algorithm for the DGP which has complexity \(O(2^K n^2)\) is exponential for the DGP, but polynomial for the DGP\(_K\) for any fixed \(K\).

**Why the alias MDGP?**

The DGP\(_3\) has a special name because the determination of molecular structures from NMR distance data provided a lot of motivation for the study of distance geometry.

**EDMCP and PSDCP**

A *predistance* matrix is symmetric and has 0 diagonal.

Border a predistance matrix \(D = (D_{ij})\) with left column and top row \(D_{00} = 0, \forall 0 \leq k \leq n\) \(D_{k0} = D_{0k} = (2n \sum_j D_{kj} - \sum_{i,j} D_{ij})/(2n^2)\)

and define the matrix \(A = (A_{ij})\) where \(A_{ij} = \frac{1}{2}(D_{0i} + D_{0j} - D_{ij})\). By [Schoenberg 35] \(D\) is an EDM iff \(A\) is PSD.
Complexity

DGP:
- DGP_1: \textbf{NP}-complete [Saxe 79]
- DGP_K: \textbf{NP}-hard [Saxe 79]
- DGP: \textbf{NP}-hard (by restriction: DGP_1 \subset DGP)
- MDGP: linear on complete graphs [Dong, Wu 02]
- DGP_K: polynomial on K-trilateration graphs [Eren et al. 04]

\[ \exists \text{ order on } V \text{ such that } \forall v \in V \exists \geq K + 1 \text{ adjacent predecessors} \]

EDMCP:
- Complexity status unknown [Laurent 09]
- Polynomial cases [Pardalos et al. 03]
- For YES instances, approximate certificates to any desired accuracy can be found in polynomial time using SDP solution technology [Alfaikh et al. 99]
**Thm.**

\( \text{DGP}_1 \) with \( d : E \to \mathbb{Z}_+ \) is NP-complete

**Proof**

1. **(DGP in NP)** If \( K = 1 \) and \( d : E \to \mathbb{Z}_+ \) then \( x \in \mathbb{Z}^{\mid V\mid} \) so conditions

   \( \forall \{i, j\} \in E \mid \|x_i - x_j\| = d_{ij} \) can be checked in polynomial time

2. **(Reduction from SUBSET-SUM, i.e. given \( a \in \mathbb{Z}^n \), find \( I \subset \{0, \ldots, n-1\} \)**

   s.t. \( \sum_{i \in I} a_i = \sum_{i \not\in I} a_i \)

3. **(instance(S)→instance(D)).** Let \( V = \{v_0, \ldots, v_{n-1}\}, \)

   \( E = \{(v_i, v_{i+1} \mod n) \mid i < n\}, \forall i < n \) \( d_{i,i+1} \mod n = a_i \) and

   \( G = (V, E, d). \)

4. **(solution(D)↔solution(S)).** \( x \) is an embedding of \( G \) in \( \mathbb{R} \) iff

   \( I = \{i \mid x_i < x_{i+1} \mod n\}, \quad \neg I = \{i \mid x_i > x_{i+1} \mod n\} \)

   is a partition of the components of \( a \) satisfying SUBSET-SUM:

   \[
   \sum_I a_i - \sum_{\neg I} a_i = \sum_I (x_{i+1} - x_i) - \sum_{\neg I} (x_i - x_{i+1}) = \\
   \sum_i x_{i+1} - \sum_i x_i = 0 \text{ (because of } \mod n \text{ index sum).}
   \]
Idea of proof

SUBSET-SUM instance: $a = (2, 1, 4, 1, 2)$
Certificate: $I = \{2, 3\}$ as $x_2 < x_3$ and $x_3 < x_4$
Verification: $a_2 + a_3 = 5 = a_1 + a_4 + a_5$
DGP\textsubscript{\(K\)} complexity [Saxe ’79]

**Thm.**

For all \(K > 1\), DGP\textsubscript{\(K\)} with \(d : E \rightarrow \mathbb{Z}_+\) is NP-hard

**Proof**

1. Polynomialsly reduce 3-SAT to DGP\textsubscript{1} with \(\text{Im}(d) = \{1, 2, 3, 4\}\) (hard)

2. Polynomialsly reduce DGP\textsubscript{1} with \(\text{Im}(d) = \{1, 2, 3, 4\}\) to DGP\textsubscript{1} with \(\text{Im}(d) = \{1, 2\}\): replace edges \(e\) with \(d(e) = 3\) by \(\begin{tikzpicture}
  
  \node (1) at (0,0) {1};
  \node (2) at (1,0) {2};
  \node (3) at (0.5,0.75) {3};
  \node (4) at (0.5,-0.75) {4};
  
  \draw (1) -- (2);
  \draw (1) -- (3);
  \draw (1) -- (4);
  \draw (2) -- (3);
  \draw (2) -- (4);
  \draw (3) -- (4);
\end{tikzpicture}\) and those with \(d(e) = 4\) by \(\begin{tikzpicture}
  
  \node (1) at (0,0) {1};
  \node (2) at (1,0) {2};
  \node (3) at (0.5,0.75) {3};
  \node (4) at (0.5,-0.75) {4};
  
  \draw (1) -- (2);
  \draw (1) -- (3);
  \draw (1) -- (4);
  \draw (2) -- (3);
  \draw (2) -- (4);
  \draw (3) -- (4);
\end{tikzpicture}\).

3. Polynomialsly reduce DGP\textsubscript{1} with \(\text{Im}(d) = \{1, 2\}\) to DGP\textsubscript{\(K\)} with \(\text{Im}(d) = \{1, 2\}\): given a graph \(G\), replace edges \(e\) with \(d(e) = 1\) by \(\begin{tikzpicture}
  
  \node (1) at (0,0) {1};
  \node (2) at (1,0) {2};
  \node (3) at (0.5,0.75) {3};
  \node (4) at (0.5,-0.75) {4};
  
  \draw (1) -- (2);
  \draw (1) -- (3);
  \draw (1) -- (4);
  \draw (2) -- (3);
  \draw (2) -- (4);
  \draw (3) -- (4);
\end{tikzpicture}\) and those with \(d(e) = 2\) by \(\begin{tikzpicture}
  
  \node (1) at (0,0) {1};
  \node (2) at (1,0) {2};
  \node (3) at (0.5,0.75) {3};
  \node (4) at (0.5,-0.75) {4};
  
  \draw (1) -- (2);
  \draw (1) -- (3);
  \draw (1) -- (4);
  \draw (2) -- (3);
  \draw (2) -- (4);
  \draw (3) -- (4);
\end{tikzpicture}\) to obtain a graph \(H\) with \(\text{Im}(d) = \{3, 4, 5, 8\}\) which is embeddable in \(\mathbb{R}^{K+1}\) if \(G\) is embeddable in \(\mathbb{R}^{K}\), and apply transformations as in Item 2 above to obtain a graph \(J\) with \(\text{Im}(d) = \{1, 2\}\). Result follows by induction.
Is \( DGP_K \) in \( \text{NP} \)?

- For \( K = 1 \), yes
- For \( K > 1 \), the YES certificates (i.e. the valid embeddings \( x \in X \)) may involve real numbers even for \( \text{Im}(d) \in \mathbb{Q} \) or \( \mathbb{N} \)
- Real numbers cannot be represented finitely, in general
- However, \( x \) must satisfy quadratic multivariate polynomial system
  \[
  \forall \{u, v\} \in E \quad \|x_u - x_v\|^2 = d_{uv}^2
  \]
- Components of \( x \) are algebraic numbers, which can be represented finitely
- **Open problem:** Find a finite representation of algebraic numbers which makes it possible to store \( x \) in space polynomial in the instance size
Search in continuous space
Continuous formulation

Solving the system

$$\forall \{i, j\} \in E \quad ||x_i - x_j|| = d_{ij}, \quad (1)$$

is numerically challenging

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

$\Rightarrow$ square both sides

- Usually, cast as a penalty objective to be minimized

$$\min_{x} \sum_{\{i, j\} \in E} (||x_i - x_j||^2 - d_{ij}^2)^2. \quad (2)$$

- 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

### Results

<table>
<thead>
<tr>
<th>Atoms</th>
<th>Variables</th>
<th>sBB</th>
<th>VNS</th>
<th>MLSL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>OF Value</td>
<td>Time</td>
<td>OF Value</td>
</tr>
<tr>
<td>cube8</td>
<td>24</td>
<td>0</td>
<td>0.22</td>
<td>0</td>
</tr>
<tr>
<td>cube27</td>
<td>81</td>
<td>0</td>
<td>30.39</td>
<td>0</td>
</tr>
<tr>
<td>cube64</td>
<td>192</td>
<td>0</td>
<td>2237.73</td>
<td>0</td>
</tr>
<tr>
<td>lavor5</td>
<td>15</td>
<td>0</td>
<td>0.02</td>
<td>0</td>
</tr>
<tr>
<td>lavor10</td>
<td>30</td>
<td>0</td>
<td>1.12</td>
<td>0</td>
</tr>
<tr>
<td>lavor20</td>
<td>60</td>
<td>0</td>
<td>2.25</td>
<td>0</td>
</tr>
<tr>
<td>lavor30</td>
<td>90</td>
<td>0</td>
<td>488.87</td>
<td>0</td>
</tr>
<tr>
<td>lavor40</td>
<td>120</td>
<td>-</td>
<td>-</td>
<td>0.09</td>
</tr>
<tr>
<td>lavor50</td>
<td>150</td>
<td>-</td>
<td>-</td>
<td>0</td>
</tr>
</tbody>
</table>
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]
\[ \langle f \rangle_{\lambda}(x) = \frac{1}{\pi^{n/2} \lambda^n} \int_{\mathbb{R}^n} f(y) e^{-\frac{||y-x||^2}{\lambda^2}} \, dy \]
Smoothing-based methods

- **dgsol:**
  
  trace curve $\bar{x}(\lambda)$ of local minimizers of $\langle f \rangle_\lambda(x)$ for $\lambda \to 0$,
  
  use trust region for local NLP with previous loc. min. as starting point

- **DVS (two-phase VNS-based heuristic):**
  
  use VNS to solve $\min \langle f \rangle_\lambda(x)$, use result to restrict bounds,
  
  use VNS to solve $\min f$ subject to restricted bounds

- **DC optimization:**
  
  like dgsol but replace trust region with the DCA (primal-dual subgradient method for DC functions)

- **Hyperbolic smoothing:**
  
  different smoothing operator based on hyperbolic functional forms
Alternating Projections Algorithm

This method solves the EDMCP. Input: predistance matrix $D$

In general, APA finds a point in the intersection of two convex sets $M, M'$ if $M \cap M' \neq \emptyset$ (in infinite time)

Let $P = I - \frac{1}{n} \mathbf{1}\mathbf{1}^T$ and $A = \frac{1}{2} P(-D)P$ (Schoenberg’s EDM-PSD equivalence);

$D$ EDM iff $A \succeq 0 \iff P(-D)P \succeq 0$,

Let $D = U\Lambda U$: spectral decomposition of $D$, $\Lambda^-$ the nonpositive part, and $Q(D) = PU\Lambda^-UP$

Let $Q'(D) = D - \text{diag}(D)$

Alternate projections using $Q$ on $M = \text{negative semidefinite matrices}$ on $\{x \mid x^T\mathbf{1} = 0\}$ and $Q'$ on $M' = \text{zero diagonal matrices}$
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 $v \in V \setminus U$ s.t. $U \subseteq \delta(v)$

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

$$
\begin{align*}
\|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{align*}
\Rightarrow
\begin{align*}
\|x_v\|^2 - 2x_v \cdot x_1 + \|x_1\|^2 &= d_{1v}^2 (3) \\
\|x_v\|^2 - 2x_v \cdot x_2 + \|x_2\|^2 &= d_{1v}^2 (4) \\
\|x_v\|^2 - 2x_v \cdot x_3 + \|x_3\|^2 &= d_{1v}^2 (5) \\
\|x_v\|^2 - 2x_v \cdot x_4 + \|x_4\|^2 &= d_{1v}^2 (6)
\end{align*}

\begin{align*}
(6)-(3) \Rightarrow \\
(6)-(4) \Rightarrow \\
(6)-(5) \Rightarrow \\
x_v = \begin{pmatrix}
2(x_1 - x_4) \\
2(x_2 - x_4) \\
2(x_3 - x_4)
\end{pmatrix}
\begin{pmatrix}
\|x_1\|^2 - \|x_4\|^2 \\
\|x_2\|^2 - \|x_4\|^2 \\
\|x_3\|^2 - \|x_4\|^2
\end{pmatrix}
- \begin{pmatrix}
d_{1v}^2 - d_{4v}^2 \\
d_{2v}^2 - d_{4v}^2 \\
d_{3v}^2 - d_{4v}^2
\end{pmatrix}
\end{align*}

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
[Williams et al. 01]

Heuristic based on local changes to each atom

Let \( C = \{C_1, \ldots, C_\ell\} \) be a covering of \( V \) (decided according to chemical properties of atomic groups)

\[
\text{for } i \leq \ell \text{ do} \\
\quad \text{while } \text{termination condition not met do} \\
\quad \quad \text{for } v \in C_i \text{ do} \\
\quad \quad \quad \text{find search direction } d_v \text{ for } x_v \text{ using local NLP} \\
\quad \quad \quad \text{determine step } s_v \text{ minimizing constraint infeasibility} \\
\quad \quad \quad x_v \leftarrow x_v + s_v d_v \\
\quad \quad \text{end for} \\
\quad \text{end while} \\
\text{end for}
\]
Monotonic Basin Hopping

[Grosso et al. 09]

Let $\mathcal{L}$ be the set of local optima of

$$f(x) = \sum_{\{u,v\} \in E} (\|x_u - x_v\|^2 - d_{uv}^2)^2$$

and $\mathcal{N} : \mathbb{R}^3 \rightarrow \mathcal{P}(\mathbb{R}^3)$ be a neighbourhood structure.

Partial order $\sqsubseteq$ on $\mathcal{L}$: $x \sqsubseteq y$ implies $f(x) > f(y)$ and $y \in \mathcal{N}(x)$

**Funnel:** $\mathcal{F} \subseteq \mathcal{L}$ s.t. $\forall x \in \mathcal{F} \exists$ chain $x = x^0 \sqsubseteq x^1 \sqsubseteq \cdots \sqsubseteq x^t = \min \mathcal{F}$

---

**Algorithm:**

$$(x \in \mathcal{F}) \rightarrow \mathcal{N}(y \in \mathcal{N}(x)) \rightarrow \text{localNLP}(x^1 \in \mathcal{F}) \text{ and repeat}$$
Self-organization heuristic

[Xu et al. 03]

while termination condition not met do
Pick \{u, v\} ∈ E (\|x_u - x_v\| \neq d_{uv})
Update \(\lambda\)
Let \(x_u \leftarrow x_u + \lambda(x_u - x_v)\)
Let \(x_v \leftarrow x_v + \lambda(x_v - x_u)\)
end while

\(\lambda\) depends on \(u, v, d\) and a given probability factor
Let \( X = (x_{ij}) \) be \( K \times |V| \); then \( \|x_i - x_j\|^2 = e_{ij}^T X^T X e_{ij} \), where \( e_{ij} \in \{-1, 0, 1\}^{|V|} \) s.t.:

\[
e_{ij\ell} = \begin{cases} 
1 & \ell = i \\
-1 & \ell = j \\
0 & \text{othw}
\end{cases}
\]  

Formulate DGP as:

\[
\forall \{i, j\} \in E \quad e_{ij}^T X^T X e_{ij} = d_{ij}^2 \\
Y = X^T X
\]

Relaxation: \( Y = X^T X \Rightarrow Y \succeq X^T X \)
Enters rigidity: ABBIE

[Hendrickson 95]

- Find largest uniquely realizable rigid components of $G$
- Contract them to obtain a graph minor $G'$ of $G$
- Repeat until (finite) convergence
- Arrange components in space using a multistart GO heuristic
Vertex orders
Orders and discretization

For each $v \in V$ let $\gamma(v) = \{u \in V \mid u < v\}$
For each $v \in V$ let $\delta(v) = \{u \in V \mid \{u, v\} \in E\}$
$\alpha(v) = \gamma(v) \cap \delta(v) =$ adjacent predecessors of $v$

$K$-trilateration order: $\forall v \in V \ (|\alpha(v)| \geq K + 1)$

Thm.

With prob. 1, yields polynomial alg. for embedding in $\mathbb{R}^L, \forall L \leq K$

Proof

Geometric build-up system is $K \times K$ and invertible with prob. 1

…but not enough to embed in $\mathbb{R}^{K+1}$

A $(K - 1)$-trilateration order guarantees rigidity with probability 1, but not uniqueness

Example: 1-trilateration in $\mathbb{R}^2$
2-trilateration in $\mathbb{R}^3$

- Let $v \in V$ and suppose $|\alpha(v)| = 3$
- $v$ belongs to the intersection $S$ of the 3 spheres centered at the adjacent predecessors
- In general, $S$ has cardinality 0, 1 or 2

“In general” means: as long as $d$ obeys strict triangular inequality on triangles $\{u_3, u_2, u_1\}$, otherwise $|S| = 2^{\aleph_0}$

CERFACS Seminar, 9 Dec. 2010 – p. 35
For $v \in V$ with $|\alpha(v)| = K$, pick a $w \in \alpha(v)$; then

$\forall u \in \alpha(v) \setminus \{w\}$

$2(x_u - x_w) \cdot x_v = (\|x_u\|^2 - d_{uw}^2) - (\|x_w\|^2 - d_{wv}^2)$  \hspace{1cm} (8)

$\|x_v\|^2 - 2x_w \cdot x_v + \|x_w\|^2 = d_{wv}^2$  \hspace{1cm} (9)

Suppose all predecessors of $v$ already embedded; then (8) is a $(K - 1) \times K$ linear system $Ax = b$, and (9) is a single quadratic equation.

If $\text{rank}(8) = K - 1$, write (8) as $x_v^{K-1} = B^{-1}(b - N x_v^K)$ where $x_v^{K-1}$ is the vector of the first $K - 1$ components of $x_v$ and $B, N$ is a basic/nonbasic partition of $A$.

Replace $x_v^{K-1}$ in (9), obtaining a quadratic equation in $x_v^K$: yields two distinct vectors $x_v, x'_v$ with probability 1.

If $\text{rank}(8) \leq K - 1$ then there are infinitely many solutions: this corresponds to the simplex $\mathcal{U}_v = \text{conv}\{x_u \mid u \in \alpha(v)\}$ having 0 volume.
Simplex volume

Condition \( \text{Vol}(\mathcal{U}_v) > 0 \) can be translated on a condition on the distance function \( d \); let

\[
\Delta_{K-1}(\mathcal{U}_v) = \frac{(-1)^K}{2^{K-1}((K-1)!)^2} \begin{vmatrix}
0 & 1 & 1 & \ldots & 1 \\
1 & 0 & d_{12}^2 & \ldots & d_{1K}^2 \\
1 & d_{12}^2 & 0 & \ldots & d_{2K}^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & d_{1K}^2 & d_{2K}^2 & \ldots & 0
\end{vmatrix}
\] (10)

where \( d_{ij} \) is the distance between \( x_{u_i} \) and \( x_{u_j} \), with \( \alpha(v) = \{u_1, \ldots, u_K\} \)

Since \( \text{Vol}(\mathcal{U}_v) = \Delta_{K-1}(\mathcal{U}_v) \), \( \Delta_{K-1}(\mathcal{U}_v) > 0 \) imply \( \text{rank}(8) = K - 1 \), i.e. one or two embeddings for \( x_v \)

\( \Delta_{K-1}(\mathcal{U}_v) > 0 \) are called strict simplex inequalities
Conditions for a discrete procedure

Having an order on $V$ such that:

1. Each vertex has at least $K$ adjacent predecessors
2. At least $K$ such pred. define a nonzero volume in $\mathbb{R}^{K-1}$

Consider the subset $\mathcal{D}$ of DGP$_K$ instances for which there is an order s.t.

1. $\forall v \in V \exists U_v \subseteq \alpha(v) (|U_v| = K)$
2. $\Delta_{K-1}(U_v) > 0$

Can we recognize whether a DGP$_K$ is in this subset or not?

Not aprioristically: verifying whether $\Delta_{K-1}(U_v) > 0$ requires finding an embedding for $U_v$

Address the issue by requiring that $G[U_v] = K_K$, the clique on $K$ vertices

This means that $d_{ij}$ is available in the problem input for all $i, j \in U_v$
Remarks on $G[U_v] = K_K$

- Strong condition: some instances might satisfy
  $\forall v \in V \exists U_v \subseteq \alpha(v) (|U_v| = K) \land \Delta_{K-1}(U_v) > 0$ but not $G[U_v] = K_K$

- In practice, the absence of $G[U_v] = K_K$ is not a problem, for an iterative procedure will establish whether a problem is in $D$ at the same time as producing a solution for it

- Condition required from a formal point of view in order to study computational complexity

- Consider the subset $D'$ of $\text{DGP}_K$ such that there is an order on $V$ with $\forall v \in V \exists U_v \subseteq \alpha(v) (|U_v| = K)$

- Because the zero-volume simplices have Lebesgue measure zero in the set of all simplices in $\mathbb{R}^K$, $D' \setminus D$ has measure zero

- Might consider dropping strict simplex inequalities altogether
Finding a good vertex order


**Discretizing Vertex Order Problem (DVOP).** Given an undirected graph $G = (V, E)$ and a positive integer $K$, establish whether there is an order $<$ on $V$ such that:

- the first $K$ vertices form a clique;
- each vertex having rank greater than $K$ has at least $K$ adjacent predecessors
For DGP, \textbf{NP}-complete by reduction from $K$-clique

Polynomially solvable for DGP$_K$ (fixed $K$):

1. find initial $K$-clique $U$ in $O\left(\binom{n}{K}\right)$ (polynomial for fixed $K$)

2. order in time in $O(n \log n)$ the set $V \setminus U$ by star adjacent to $U$ and, if a draw, by star size

3. greedy on $V \setminus U$

The inherent difficulty of the DVOP is in choosing the first $K$-clique

Since $K \ll n$, not a bad situation at all!
Math. Prog. Formulation

- Sets.
  1. $G = (V, E)$: the graph (assume $V = \{1, 2, \ldots, |V|\}$)

- Parameters: $K \in \mathbb{N}$.

- Decision variables: vertex rank in $<_H$:

$$\forall v \in V, k \in V \quad y_{vk} = \begin{cases} 1 & v \text{ is } k\text{-th in the order } <_H \\ 0 & \text{otherwise.} \end{cases}$$ (11)

- Objective: none

- Constraints:
  1. each rank $k$ has a unique vertex assigned to it: $\forall k \in V \sum_{v \in V} y_{vk} = 1$;
  2. each vertex has a unique rank assigned to it: $\forall v \in V \sum_{k \leq K} y_{vk} = 1$;
  3. each vertex with rank $> K$ has at least $K$ adjacent predecessors in $<_H$:

$$\forall v \in V, k \in V \setminus \{1, \ldots, K\} \quad \sum_{u \in V} \sum_{(u, v) \in A, i < k} y_{ui} \geq Ky_{vk}.$$ (12)
Discretization
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].

The backbone is a total order $<$ on a set $V$ of atoms.
Protein distances

Assumption: known distances between consecutive atoms

Assumption: known angles between atom triplets

Hence $d_{i,i+2}$ are also known for all $i < |V| - 1$

Assumption: distances $d_{i,i+3}$ always $< 6 \text{Å}$, so they can be measured using NMR techniques

In general, we do not know distances $d_{i,i+4}$

⇒ We know distances to at least 3 adjacent predecessors, so backbone order $< \text{is a 2-trilateration order on } V$

but not a 3-trilateration order, cannot use poly alg.
Algorithmic idea for solving backbone placement problem:

1. place first 3 vertices and obtain partial embedding $x$, set $j = 4$

2. at atom with rank $j$, solve the system

$$\forall i \in \{j - 3, j - 2, j - 1\} \quad \|x_i - x_j\| = d_{ij} \quad (13)$$

3. for each position found, extend $x$ to atom $j$, increase $j$ and repeat from Step 2

Since the search is branched at Step 3, we obtain a search tree with at most two subnodes per node.
Using other distances

In order for the above algorithm to work, we are only using distances from $j$ to $j-1$, $j-2$, $j-3$

When proteins coil, backbone atoms of very distant ranks might be placed very close

If atoms $i,j$ with $|i-j| >> 4$ are placed closer than 5Å, then NMR finds $d_{ij}$

Such distances must be used to prune out invalid embeddings: replace (13) by $\forall i \in \alpha(j) \|x_i - x_j\| = d_{ij}$
The Branch-and-Prune algorithm

\( j \): rank of current atom \( x < j \): partial embedding to rank \( j - 1 \)

\( G \): instance \( \hat{X} \): current pool of embeddings

\( S(y, r) \): 3-sphere centered at \( y \) with radius \( r \)

\( \text{BRANCHANDPRUNE}(j, x < j, G, \hat{X}) \):

Let \( S \leftarrow \bigcap_{k \leq 3} S(x_{j-k}, d_{j-k,j}) = \{s_1, \ldots, s_q \} \), where \( q \in \{0, 1, 2\} \)

\( \text{for } p \leq q \text{ do} \)

Extend the current embedding to \( x^{(p)} = (x < j, s_p) \)

if \( \forall i \in \alpha(v) \|x_i^{(p)} - x_j^{(p)}\| = d_{ij} \) then

if (\( j = n \)) then

Let \( \hat{X} \leftarrow \hat{X} \cup \{x^{(p)}\} \)

else

\( \text{BRANCHANDPRUNE}(j + 1, x^{(p)}, G, \hat{X}) \)

end if

end if

end for
Discretizable MDGP

[Lavor et al., COAP, in revision]

\( U_0 \): the set of the first \( K \) vertices  
\( U_v \): \( v \) and its \( K \) immediate predecessors

**Defn.**

**Discretizable MDGP (DMDGP)** Given a weighted undirected graph \( G = (V, E, d) \) where \( d : E \to \mathbb{R}_+ \) and a total order < on \( V \) such that:

1. \( G[U_0] \) is the clique on \( K \) vertices (START);

2. for all \( v \) s.t. \( \rho(v) > K \), \( G[U_v] \) is the \( (K + 1) \)-clique (DISCRETIZATION);

3. for all triplets \( (u, v, z) \) of consecutive vertices we have \( d_{uz} < d_{uv} + d_{vz} \) (STRICT TRIANGULAR INEQUALITY),

find an embedding \( x : V \to \mathbb{R}^K \) such that \( \|x_i - x_j\| = d_{ij} \) for all \( \{i, j\} \in E \)

**Thm.**

BP finds all non-isometric embeddings of \( G \)
Complexity

- The BP algorithm is an exponential-time algorithm in the worst case.
- The DMDGP is \textbf{NP}-hard (reduction from \textsc{Subset-Sum}).
- The DMDGP is the “smallest” \textbf{NP}-hard problem of its type, in the sense that if $|U_v| = K + 2$ instead of $K + 1$ for each $v$, then the order is a $K$-trilateration order and the instance can be solved polynomially.

\begin{center}
\textbf{In practice, it beats the hell out of its competitors!}
\end{center}

\ldots both for speed and accuracy.
### Comparison with *dgsol* 1/4

| Instance | $|V|$ | $|S|$ | **dgsol** | **BP-One** | **BP-All** |
|----------|-----|----|----------|------------|------------|
| Name     |     |    | CPU      | LDE        | CPU        | LDE        | CPU  | #Sol |
| mmorewu-2 | 8   | 28 | 0.02     | 2.63E+5    | 0.00       | 4.37E-10   | 0.00 | 2    |
| mmorewu-3 | 27  | 331| 0.23     | 6.99       | 0.00       | 2.97E-09   | 0.00 | 2    |
| mmorewu-4 | 64  | 1882| 0.67    | 7.79E-6    | 0.00       | 5.56E-09   | 0.00 | 4    |
| mmorewu-5 | 125 | 7105| 2.94     | 3.54E-6    | 0.00       | 1.67E-08   | 0.01 | 4    |
| mmorewu-6 | 216 | 21461| 18.65  | 0.032      | 0.02       | 4.91E-08   | 0.03 | 4    |

CERFACS Seminar, 9 Dec. 2010 – p. 51
Comparison with \texttt{dgsol} 2/4

| Instance | $|V|$ | $|S|$ | \texttt{dgsol} | \texttt{BP-One} | \texttt{BP-All} |
|----------|-----|-----|------------|------------|-------------|
| lavor10_0 | 10 | 33 | 0.02 | 1.57E-5 | 0.00 | 5.36E-10 | 0.00 | 4 |
| lavor15_0 | 15 | 57 | 0.10 | 4.04E-5 | 0.00 | 2.84E-09 | 0.00 | 16 |
| lavor20_0 | 20 | 105 | 0.14 | 2.77E-5 | 0.00 | 6.13E-09 | 0.00 | 8 |
| lavor25_0 | 25 | 131 | 0.84 | 1.18E-4 | 0.00 | 1.38E-09 | 0.00 | 8 |
| lavor30_0 | 30 | 169 | 0.40 | 1.75E-5 | 0.00 | 1.23E-09 | 0.00 | 2 |
| lavor35_0 | 35 | 171 | 0.81 | 9.33E-5 | 0.00 | 1.52E-09 | 0.00 | 64 |
| lavor40_0 | 40 | 295 | 2.84 | 0.096 | 0.00 | 2.87E-09 | 0.00 | 2 |
| lavor45_0 | 45 | 239 | 3.33 | 0.170 | 0.00 | 6.92E-09 | 0.00 | 2 |
| lavor50_0 | 50 | 271 | 3.45 | 0.696 | 0.00 | 3.96E-08 | 0.46 | 4096 |
| lavor55_0 | 55 | 551 | 5.80 | 0.257 | 0.00 | 2.66E-09 | 0.00 | 64 |
| lavor60_0 | 60 | 377 | 5.15 | 0.049 | 0.00 | 3.51E-09 | 0.00 | 64 |
| lavor65_0 | 65 | 267 | 2.61 | 0.065 | 0.00 | 7.76E-10 | – | – |
| lavor70_0 | 70 | 431 | 8.73 | 0.107 | 0.02 | 1.64E-09 | – | – |
## Comparison with \texttt{dgsol} 3/4

| Name      | \(|V|\) | \(|S|\) | \texttt{dgsol} | \texttt{BP-One} | \texttt{BP-All} |
|-----------|--------|--------|---------------|-----------------|----------------|
| \texttt{lavor100.2} | 100    | 605    | 6.95 0.167    | 2.26 4.01E-09   | – –           |
| \texttt{lavor200.2} | 200    | 1844   | 63.52 0.395   | 0.00 5.66E-08   | – –           |
| \texttt{lavor300.2} | 300    | 2505   | 100.99 0.261  | 0.03 1.56E-08   | – –           |
| \texttt{lavor400.2} | 400    | 2600   | 182.21 0.767  | 0.01 3.35E-09   | – –           |
| \texttt{lavor500.2} | 500    | 4577   | 329.29 0.830  | 0.27 4.69E-07   | – –           |
| \texttt{lavor600.2} | 600    | 5473   | 299.76 0.700  | 0.01 4.94E-08   | – –           |
| \texttt{lavor700.2} | 700    | 4188   | 281.34 0.569  | 0.16 1.83E-06   | – –           |
| \texttt{lavor800.2} | 800    | 6850   | 570.20 0.528  | 3.34 3.37E-06   | – –           |
| \texttt{lavor900.2} | 900    | 7965   | 550.26 0.549  | 3.08 5.62E-06   | – –           |
| \texttt{lavor1000.2} | 1000   | 8229   | 844.52 0.695  | 0.81 2.04E-06   | – –           |
## BP performance

<table>
<thead>
<tr>
<th>Name</th>
<th></th>
<th>V</th>
<th>E</th>
<th>CPU</th>
<th>LDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lbrv</td>
<td></td>
<td>57</td>
<td>476</td>
<td>0.00</td>
<td>1.54e-14</td>
</tr>
<tr>
<td>laqr</td>
<td></td>
<td>120</td>
<td>929</td>
<td>0.00</td>
<td>1.86e-09</td>
</tr>
<tr>
<td>2erl</td>
<td></td>
<td>120</td>
<td>1136</td>
<td>0.00</td>
<td>1.33e-14</td>
</tr>
<tr>
<td>lcrn</td>
<td></td>
<td>138</td>
<td>1250</td>
<td>0.00</td>
<td>2.24e-13</td>
</tr>
<tr>
<td>lahl</td>
<td></td>
<td>147</td>
<td>1205</td>
<td>0.00</td>
<td>1.50e-09</td>
</tr>
<tr>
<td>lptq</td>
<td></td>
<td>150</td>
<td>1263</td>
<td>0.00</td>
<td>2.30e-13</td>
</tr>
<tr>
<td>lbrz</td>
<td></td>
<td>159</td>
<td>1394</td>
<td>0.00</td>
<td>3.53e-13</td>
</tr>
<tr>
<td>lhoe</td>
<td></td>
<td>222</td>
<td>1995</td>
<td>0.00</td>
<td>3.18e-13</td>
</tr>
<tr>
<td>11fb</td>
<td></td>
<td>232</td>
<td>2137</td>
<td>0.00</td>
<td>5.31e-14</td>
</tr>
<tr>
<td>lptf</td>
<td></td>
<td>249</td>
<td>2283</td>
<td>0.00</td>
<td>2.73e-12</td>
</tr>
<tr>
<td>1jk2</td>
<td></td>
<td>270</td>
<td>2574</td>
<td>0.00</td>
<td>2.09e-13</td>
</tr>
<tr>
<td>l39a</td>
<td></td>
<td>303</td>
<td>2660</td>
<td>0.00</td>
<td>2.68e-12</td>
</tr>
<tr>
<td>lacz</td>
<td></td>
<td>324</td>
<td>3060</td>
<td>0.00</td>
<td>3.15e-12</td>
</tr>
<tr>
<td>lpoa</td>
<td></td>
<td>354</td>
<td>3193</td>
<td>0.00</td>
<td>1.36e-13</td>
</tr>
<tr>
<td>lfs3</td>
<td></td>
<td>378</td>
<td>3443</td>
<td>0.00</td>
<td>8.08e-13</td>
</tr>
<tr>
<td>lmbn</td>
<td></td>
<td>459</td>
<td>4599</td>
<td>0.00</td>
<td>1.36e-09</td>
</tr>
<tr>
<td>lrgs</td>
<td></td>
<td>792</td>
<td>7626</td>
<td>0.00</td>
<td>4.22e-13</td>
</tr>
<tr>
<td>l40</td>
<td></td>
<td>1224</td>
<td>20382</td>
<td>0.02</td>
<td>1.00e-12</td>
</tr>
<tr>
<td>lpm</td>
<td></td>
<td>1443</td>
<td>14292</td>
<td>0.02</td>
<td>2.85e-13</td>
</tr>
<tr>
<td>n4w</td>
<td></td>
<td>1610</td>
<td>16940</td>
<td>0.02</td>
<td>1.19e-12</td>
</tr>
<tr>
<td>lmq</td>
<td></td>
<td>2032</td>
<td>19564</td>
<td>0.02</td>
<td>4.90e-12</td>
</tr>
<tr>
<td>lrwh</td>
<td></td>
<td>2265</td>
<td>21666</td>
<td>0.02</td>
<td>2.08e-13</td>
</tr>
<tr>
<td>3b34</td>
<td></td>
<td>2790</td>
<td>29188</td>
<td>0.07</td>
<td>1.17e-11</td>
</tr>
<tr>
<td>2e7z</td>
<td></td>
<td>2907</td>
<td>42098</td>
<td>0.08</td>
<td>4.26e-12</td>
</tr>
<tr>
<td>lep</td>
<td></td>
<td>3861</td>
<td>35028</td>
<td>0.16</td>
<td>3.19e-12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Name</th>
<th>CPU</th>
<th>LDE</th>
</tr>
</thead>
<tbody>
<tr>
<td>lbrv</td>
<td>1.48</td>
<td>2.74e-01</td>
</tr>
<tr>
<td>laqr</td>
<td>7.77</td>
<td>4.88e-01</td>
</tr>
<tr>
<td>2erl</td>
<td>9.38</td>
<td>2.92e-01</td>
</tr>
<tr>
<td>lcrn</td>
<td>9.47</td>
<td>2.24e-01</td>
</tr>
<tr>
<td>lahl</td>
<td>6.95</td>
<td>1.46e-01</td>
</tr>
<tr>
<td>lptq</td>
<td>9.16</td>
<td>1.21e-01</td>
</tr>
<tr>
<td>lbrz</td>
<td>7.77</td>
<td>4.88e-01</td>
</tr>
<tr>
<td>lhoe</td>
<td>16.83</td>
<td>2.06e-01</td>
</tr>
<tr>
<td>11fb</td>
<td>38.94</td>
<td>2.88e-01</td>
</tr>
<tr>
<td>lptf</td>
<td>42.50</td>
<td>2.00e-01</td>
</tr>
<tr>
<td>1jk2</td>
<td>86.98</td>
<td>4.05e-01</td>
</tr>
<tr>
<td>l39a</td>
<td>37.24</td>
<td>2.80e-01</td>
</tr>
<tr>
<td>lacz</td>
<td>35.97</td>
<td>3.97e-01</td>
</tr>
<tr>
<td>lpoa</td>
<td>64.03</td>
<td>4.67e-01</td>
</tr>
<tr>
<td>lfs3</td>
<td>54.68</td>
<td>2.69e-01</td>
</tr>
<tr>
<td>lmbn</td>
<td>124.24</td>
<td>4.46e-01</td>
</tr>
<tr>
<td>lrgs</td>
<td>237.93</td>
<td>4.69e-01</td>
</tr>
<tr>
<td>l40</td>
<td>1142.49</td>
<td>4.89e-01</td>
</tr>
<tr>
<td>lpm</td>
<td>398.29</td>
<td>5.06e-01</td>
</tr>
<tr>
<td>n4w</td>
<td>994.51</td>
<td>5.26e-01</td>
</tr>
<tr>
<td>lmq</td>
<td>451.58</td>
<td>5.40e-01</td>
</tr>
<tr>
<td>lrwh</td>
<td>934.29</td>
<td>5.38e-01</td>
</tr>
<tr>
<td>3b34</td>
<td>940.95</td>
<td>6.47e-01</td>
</tr>
<tr>
<td>2e7z</td>
<td>915.39</td>
<td>6.40e-01</td>
</tr>
<tr>
<td>lep</td>
<td>2037.86</td>
<td>4.92e-01</td>
</tr>
</tbody>
</table>
Remarks and extensions

Remarks:
- Cardinality of solution set is always a power of two
- Looks linear-time! And the exponential increase?

Extensions:
- To the DMDGP extension to an arbitrary dimension $K$
- To DDGP (replace “immediate predecessors” with “adjacent predecessors”) and $\text{DDGP}_K$ [Mucherino et al., Opt. Lett., in revision]
- To some data limitations (H-H distances) found in protein-related NMR data [Lavor et al., JOGO, accepted; Lavor et al., IEEE 2009]
- To some data uncertainty found in some protein-related NMR data [Lavor et al., IEEE 2010]
Symmetry
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.
An empirical observation

| Instance       | |X| |
|----------------|---|---|
| mmorewu-2      | 2 | 2 |
| mmorewu-3      | 2 | 1 |
| mmorewu-4      | 2 | 1 |
| mmorewu-5      | 2 | 1 |
| mmorewu-6      | 2 | 1 |
| lavor10_0      | 4 | 8 |
| lavor15_0      | 16 | 16 |
| lavor20_0      | 8 | 8 |
| lavor25_0      | 8 | 8 |
| lavor30_0      | 2 | 2 |
| lavor35_0      | 2 | 2 |
| lavor40_0      | 2 | 2 |
| lavor45_0      | 2 | 2 |
| lavor50_0      | 2 | 2 |
| lavor55_0      | 2 | 2 |
| lavor60_0      | 2 | 2 |

For all tested DMDGP instances, \( \exists \ell \in \mathbb{N} \) such that \(|X| = 2^\ell\)
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
(e.g. symmetric paths to level $|V|$ from nodes 16 and 45)
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 (e.g. symmetric paths to level $|V|$ from nodes 16 and 45)
Discretization/pruning distances

- Define arithmetic operations on $V$ according to the vertex rank in the DMDGP order:
  - $\forall u, v \in V \quad u - v = \text{rank}(u) - \text{rank}(v)$
  - $\forall u \in V, \ell \in \mathbb{N} \quad u - \ell = \text{rank}(u) - \ell$

- Let $E_D = \{\{u, v\} | |u - v| \leq K\}$ and $E_P = E \setminus E_D$

- $E_D$ 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
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 $d_{15}$
Reflection symmetry

[Liberti et al. arXiv 2010]

Given an embedding $x$, let $R^v_x$ 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.

$$\forall x \in X \ (\|x_v - x_u\| \in H^{uv})$$

Further, for $x' \in X \setminus \{x\}$,

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

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

The $g_v$’s map $x$ to its partial reflection with first branch at $v$
Discretization group

The following hold with probability 1 \( \forall u > K \):

1. \( g_v \) is injective (by reflection)
2. \( g_v \) 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 (isomorphic to \( C_{m-K}^{m-K} \))

By previous thm, discretization distances are invariant under \( \mathcal{G}_D \)

The action of \( \mathcal{G}_D \) on \( X \) is transitive, i.e. \( \forall x' \in X \exists g \in \mathcal{G}_D \ (x' = g(x)) \)

This action has only one orbit, i.e. \( X = \mathcal{G}_D x \)
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 \( w \in \{u + 4, \ldots, v\} \)

In order to keep invariance we remove such \( g_w \)'s from the group

**Pruning group:** \( G_P = \langle g_w \mid w > K \land \forall \{u, v\} \in E_P \ (w \not\in \{u + K, \ldots, v\}) \rangle \)

\( G_P \leq G_D \) and all distances are invariant w.r.t. the pruning group

Again, action of \( G_P \) on \( X \) is transitive (nontrivial proof)
**Power of two**

**Thm.**

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

**Proof**

With probability 1:

- \( \mathcal{G}_D \cong C_2^{n-K} \Rightarrow |\mathcal{G}_D| = 2^{n-K} \)
- \( \mathcal{G}_P \leq \mathcal{G}_D \Rightarrow |\mathcal{G}_P| \parallel |\mathcal{G}_D| \Rightarrow \exists \ell \in \mathbb{N} \ |\mathcal{G}_P| = 2^\ell \)
- 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
Polynomial cases
BP complexity

As mentioned above, the BP has worst-case exponential complexity.

Complexity depends on BP nodes; since height $\leq |V|$, only need to consider treewidth.

- If $\not\exists$ pruning edges $\{u, v\} \in E_p$, number of nodes at level $\ell \in \{K + 1, \ldots, |V|\}$ is $2^{\ell - K}$.

- For two pruning edges $\{u, v\}, \{u', v\}$ with $u < u'$, with prob. 1 $\{u, v\}$ prunes all nodes pruned by $\{u', v\}$ (and more).

$\Rightarrow$ For YES instances and with prob. 1, only longest edges incident to $v$ need be considered.

- 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 $u$

This row: no pruning

$BP \text{ nodes vs. pruning edges } \{u, v\}$

1st line: $v - u$

vertices: $|BP \text{ nodes}|$ at level $v$

arcs: $\exists$ pruning edge $\{u + \text{arc}_\text{label}, v\}$
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: $v_0 = K + 3$
BP complexity: $O(2^{v_0}|V|)$

Sufficient: $\exists v_0 \text{ s.t. every subsequence of } s \text{ consecutive vertices } > v_0 \text{ with no incident pruning edge is preceded by a vertex } v_s$

s.t. $\exists u_s < v_s \quad (v_s - u_s \geq |s| \land \{u_s, v_s\} \in E_P)$

“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

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

Artificial “Lavor” instances do not seem to exhibit this property (they are difficult instances)
Remarks on symmetry theory

- Both the “power of two” property and the BP speed are explained by a single theory of symmetry.

- Theory only holds for the DMDGP (also extended to $K$ dimensions), **not for the DDGP**.

- I.e. the assumption that the predecessors are *immediate* is explicitly used in proofs.

- There is a counterexample showing that *adjacent predecessors* do not yield all the reflection symmetry.

- To date, however, we know of no Lebesgue measure 1 DDGP instance set with $|X| \neq 2^\ell$. 
Attacking the real problem
Virtual hydrogen backbone

The most accurate NMR distances are between hydrogen atoms only, but the actual backbone is a chain of \(N-C_\alpha-C\) 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.
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 sequence \( r : \mathbb{N} \rightarrow V \cup \{0\} \) such that
\[
\exists \ell \in \mathbb{N} \text{ with } \forall i \leq \ell \ (r_i \in V) \land \forall i > \ell \ (r_i = 0).
\]

- 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 $[d_{uv}^L, d_{uv}^U]$

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
Are re-orders magic?

We asked ourselves whether we could have found a normal vertex order (rather than a re-order) to deal with proteins.

We haven’t run the tests yet, but we believe our DVOP techniques could find such an order; but at the moment the question is open.

If not, then simply listing the same information twice would somehow yield a new mathematical entity that could not previously be obtained:

no new information $\Rightarrow$ new results? seems unlikely...

The following, however, might be possible:

- Re-orders allow a compact, parametrized description of a discretization order for a whole class of instances.
- DVOP techniques would yield a new order for each instance — the different orders might not be easily described with a compact expression.
Implementations
Sequential code

Mucherino et al. LNCS 2010

- The code is available in open source
- Any doubt, ask the MASTER (Antonio)
**Parallel code**

**Seconds of user CPU on Grid5000** ([www.grid5000.fr](http://www.grid5000.fr))

| $|V|$  | CPUs             |
|------|------------------|
| 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 $G_P x = X$ result from symmetry to obtain all solutions from just one
- Extend complexity study to actual problem with discrete/uncertain distances
- Better understand re-orders
- Progress on “MDGP $\in \text{NP}$?” question
Conclusion

- BP: faster than all continuous methods, more accurate than most\(^1\)
- Only method that finds all non-isometric solutions
- Only method (from optimization community) that addresses the NMR limitation concerning H-H distances
- Capable of dealing with typical NMR errors found in proteins
- Supported by a comprehensive theoretical analysis concerning:
  - \textbf{NP}-hardness
  - Symmetry
  - Polynomial cases
- Comes with an open-source implementation
- Can be parallelized effectively

\(^1\) Less accurate than semidefinite facial reduction (RMSD: \(10^{-7}\) vs. \(10^{-14}\)) [Krislock 2010]