Discovery Thread: Project 1

Consider the chemical compound database assigned to your team. Read n, the integer on the first line of the datafile. Extract $X = (X(k))_{1 \leq k \leq n}$, $Y = (Y(k))_{1 \leq k \leq n}$, $Z = (Z(k))_{1 \leq k \leq n}$, $Q = (Q(k))_{1 \leq k \leq n}$ from lines $3 : n+2$, and columns 2:5. Extract also the list of atoms and create a vector of size n of characters from the set $\{'C','O','H','N','F'\}$.

1. Construct the symmetric matrix $F = (F_{k,l})_{1 \leq k,l \leq n}$ defined by

   $$F_{k,l} = \frac{|Q(k)Q(l)|}{\sqrt{(X(k)-X(l))^2 + (Y(k)-Y(l))^2 + (Z(k)-Z(l))^2}} \ , \ 1 \leq k,l \leq n \ , \ k \neq l$$

   Find a threshold $\tau > 0$ so that at least half of the entries in F are smaller than or equal to τ and half of the entries are larger than or equal to τ. Compute the weight matrix W by thresholding F:

   $$W_{k,l} = \begin{cases} F_{k,l} & \text{if} \quad F_{k,l} \geq \tau \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$$

2. Construct the graph Laplacian $\Delta = D - W$ and compute its eigenpairs.

3. Determine the plan embedding using the Graph Visualization Spectral Algorithm. Denote by $\{x_1, x_2, \ldots, x_n\} \subset \mathbf{R}^2$ the n points in plan.

4. Plot the planar embedding using circles of different colors for the atoms of different type. Draw only edges associated to strictly positive weight.

5. Extend the 2D embedding into a 3D embedding by adding a 0 on the third component of each 2D vectors determined before. Denote $\{u_1, u_2, \ldots, u_n\}$ the 3D points, where $u_k = [x_k^T \ 0]^T$.

6. Find the optimal rigid transformation that best maps the n geometric point $\{(X(k), Y(k), Z(k)) \ ; \ 1 \leq k \leq n\}$ to $\{u_1, u_2, \ldots, u_n\}$.

7. Draw on the same figure the two geometric graphs (vertices and edges).

8. Compute the modeling error:

   $$e(W) = \sum_{k=1}^{n} \| \begin{bmatrix} X_k \\ Y_k \\ Z_k \end{bmatrix} - \hat{a}\hat{Q}(u_k - \hat{z}) \|_2^2$$

9. Repeat 1-7 for an exponential potential:

   $$F_{k,l} = |Q(k)Q(l)| exp\left\{ -((X(k)-X(l))^2 + (Y(k)-Y(l))^2 + (Z(k)-Z(l))^2)/a \right\} \ , \ 1 \leq k,l \leq n$$

   where a is the average of pairwise squared distances:

   $$a = \frac{2}{n(n-1)} \sum_{1 \leq k < l \leq n} ((X(k)-X(l))^2 + (Y(k)-Y(l))^2 + (Z(k)-Z(l))^2)$$