

Assignment 7

for lecture "Bioinformatics III" WS 08/09

Return by email to p.walter@bioinformatik.uni-saarland.de until **Jan. 18**. This assignment will be discussed in the tutorial on Jan. 19, 2009, room15, building E1 3



1. Force directed layout of graphs (100 points)

Layout some graphs, first a few simple test cases, then two real interaction networks from the BIND database. Use a repulsive degree dependent Coulomb type potential

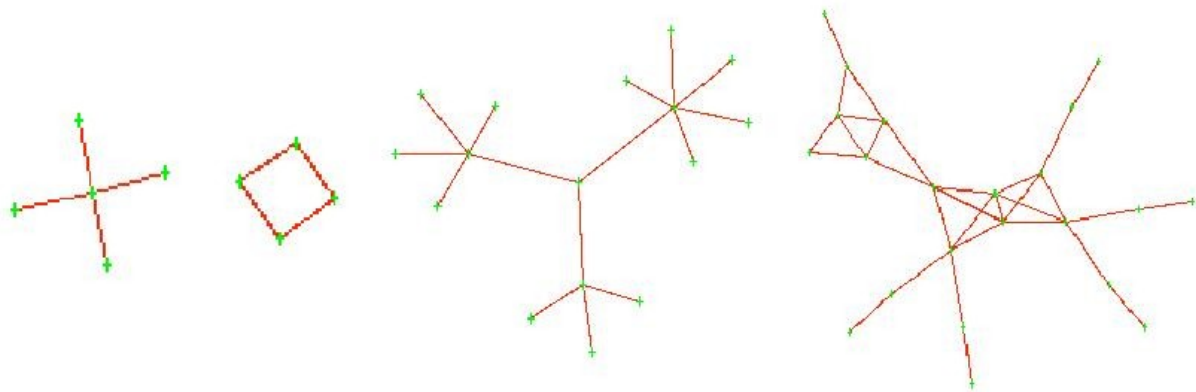
$E_c(r_{ij}) = k_i k_j / r_{ij}$ between all nodes plus a degree independent harmonic attractive potential

$E_h(r_{ij}) = r_{ij}^2 / 2$ between interacting nodes; r_{ij} is the distance between nodes i and j . Perform

this layout in 2D, therefore $r_{ij}^2 = (x_i - x_j)^2 + (y_i - y_j)^2$.

(a) Testfiles:

Start with the test files "stern.dat", "viereck.dat", "stern2.dat" and "dog.dat" (available from the teaching website). These should give you final configurations similar to the ones shown below with final energies of about 6.9, 7.3, 53.8 and 177 units. The simple ones converge after about 300 iterations, while the more complex may take up to a



thousand iterations.

For each of the networks plot the final configuration and give the final energies. Also plot the energy vs. iteration number and determine at which point the layout process can be considered converged.

What happens when you change the "friction constant" α (see step (iii) of the hint below)?

Hint: To create the layouts follow these steps — and have a look at the supplied example code:

1. Read in the interaction files and create a network from them. The files give the number of nodes on the first line and then the two endpoints of a link on each of the subsequent lines. From the network we only need the list of interactions $W[i][j] = k_i k_j$
2. Choose initial positions for all the nodes in the x - y -plane. A reasonable width is within ± 10 units from the origin.
3. For the iterated layout perform at least 500 steps for the test files and at least 5000 steps for the "real" networks:
 - i. Calculate the distances r_{ij} between the nodes and from them the resulting forces as TODO and sum them up. The total force on node i is $F_i = \sum_j F_{ij}$. Note that $F_{ij} = -F_{ji}$, the forces are symmetric.

- ii. Add a random force in the range $-0.3 \dots 0.3$ to the total force on each node. This additional “thermal” contribution improves the convergence as it helps to escape from local minima.
 - iii. Update the position of each node from the forces as $\Delta \mathbf{r}_i = \alpha \mathbf{F}_i$ with the inverse friction coefficient $\alpha = \Delta t / \gamma$. A reasonable value is $\alpha = 0.03$.
 - iv. Calculate the total energy as the sum of the individual interaction energies. TODO. Print out this energy together with the number of the iteration. You will see that the energy decreases fast in the beginning and then slower and slower.
 - v. Repeat from (i) until the total energy is essentially constant.
4. Print out and plot the positions of the nodes after the last iteration. If you use gnuplot then loop over all pairs of nodes and print out the x and y positions of the two nodes of a link on a separate line, each, followed by an empty line. Plot this file “with linespoints” to get a network representation of the layed out graph.

(b) **“Real” networks**

Now perform the same layout on the following networks. Give the final energies and configurations of each of them and when you stopped the layout process.

1. “sfnetz_100.dat” is a thinned out scale free network of 100 nodes, where every second links has been left out. When you are in time you can go for “sfnetz_200.dat”, which has twice as many nodes.
2. “11309.txt” and “2287.txt” contain networks extracted from the BIND database for the taxon identifiers 11309 and 2287, respectively (what species are these?).
3. What is the (visual) difference between "HighSociety.dat" and HighSociety_Murdered.dat"?

***Hint:** You may have to run a few trials with different initial placements of the nodes and then choose the best result.*

What happens when you skip step (ii), i.e., do the optimization without the random forces? Plot the total energy vs. iteration with and without the random forces for one of the “real” networks of your choice. Be careful to scale the axis so that the important difference can clearly be seen.

These networks contain more than a single cluster. What happens to the different clusters? Why?