

S-114.4150 Complex Networks, autumn 2006

Homework exercise, **deadline 1.1.2007**

In this homework exercise, you will analyze a network data set. Follow the instructions below, and write a short report including figures. You can return the exercise as printouts or by email (PDF or PS or .doc) to `jsaramak@lce.hut.fi`.

The data set to be analyzed depicts the collaboration network of network scientists based on two review articles in 2006, collected by M.E.J. Newman [1]. In this network, vertices depict scientists, and edges indicate that the scientists have co-authored papers.

This network is weighted - an edge between two vertices means that the corresponding scientists have coauthored at least one research paper, and the weight of this edge has been calculated from the formula

$$w_{ij} = \sum_p \frac{\delta_{i,p} \delta_{j,p}}{n_p - 1},$$

where the index p runs over all papers and n_p is the number of authors of paper p ; i.e. the weights depend on how many authors were involved in each paper. In the network to be studied in this exercise, vertices of zero degree (due to single-author papers only) have been removed, so the data set is not identical to Newman's.

The ASCII file lists all edges in the network. It contains three columns such that column 1: index of vertex i , column 2: index of vertex j , column 3: weight of edge w_{ij} . Note that each edge is listed only once; this network is undirected so the existence of an edge i, j with weight w_{ij} implies the existence of an edge j, i with the same weight, $w_{ji} = w_{ij}$.

The suggested tool for this exercise is Matlab. The network is small enough to be handled using sparse adjacency and weight matrices, so simply load the file into Matlab, count the number of vertices and edges and reserve memory: `A=spalloc(N,N,2*E)`; `W=spalloc(N,N,2*E)`. The factor 2 is due to symmetry - each edge of course appears twice ($i - j, j - i$) in the matrices. Then simply go through the edge list and fill the corresponding elements of A and W , e.g. if the first data row is `1 2 0.66`, the matrix elements should be `A(1,2)=1`; `W(1,2)=0.66`. Don't forget to symmetrize the matrices, e.g. after going through edges just write `A=A+A'`; `W=W+W'`;

Your analysis should contain at least the following:

1. The very basics: number of vertices N , number of edges E , average degree $\langle k \rangle$, average strength $\langle s \rangle$, average edge weight $\langle w \rangle$.
2. Basic unweighted characteristics: plots of the degree distribution $P(k)$ (log-binned or cumulative), average clustering coefficient as function of degree $c(k)$, average nearest-neighbour degree as function of degree $k_{nn}(k)$. Both $c(k)$ and $k_{nn}(k)$ should be displayed as double logarithmic plots (Matlab: `loglog(c,k)`). What can you say about the degree distribution? How about clustering? Is the network assortative or disassortative?

Note that due to small network size, these plots will look rather noisy. You do not have to attempt to fit anything to the data. Also note that there are vertices of degree $k = 1$, and the clustering coefficient is not defined for such vertices.

3. Is the network singly connected? If not, how large is the largest connected component? Bonus: if not singly connected, how does the component size distribution look like?

This can be done using brute force: start from a vertex, mark all of its neighbours as belonging to the same component, mark all their neighbours etc until no more neighbours found, then check if any vertex is not marked and do the same. The more elegant (and perhaps more simple way) is to use the edge percolation algorithm discussed at the last lecture - just add all edges one by one, keep track of cluster sizes and find the largest cluster when all edges have been added.

4. Basic weighted characteristics: plots of the strength distribution $P(s)$ and weight distribution $P(w)$. Again, either log-binned or cumulative, plotted on double logarithmic scale. What can you say about these?
5. Average vertex strength as function of degree, on a double logarithmic scale. (Just take all vertices of degree k and calculate their average strength, for all k . What does this curve tell you? Compare your plot to the expected form if there are no non-trivial correlations, $\langle s(k) \rangle = \langle w \rangle k$.)
6. For bonus points, do and present any analysis of your wishing. See lecture slides for hints. Some possibilities: rich-club coefficient (you do not have to compare to randomized networks, just plot as function of k or s), visualization of the largest connected component, edge percolation analysis, minimal spanning tree, ...

Some Matlab hints: degrees and strengths are readily calculated as row or column sums: `k=full(sum(A,1)); s=full(sum(W,1));` produces vectors `k` and `s` which contain strengths and degrees. The function `full` simply changes the results from sparse back to the ordinary vector format.

References

- [1] This and other data sets can be found at <http://www-personal.umich.edu/~mejn/netdata/>.