K-Nearest Neighbor

Description

The K-Nearest-Neighbor (k_{NN}) is a measure of the correlation between the degree of a node and that of its neighbors. In many systems strong correlations are observed, and one distinguishes networks into assortative and disassortative. A network is assortative if large (small) degree nodes tend to be linked with large (small) degree nodes. Social networks are typical examples of assortative networks. A network is disassortative if large (small) degree nodes tend to be linked with small (large) degree nodes. Biological and technological networks (like the Internet) are examples of disassortative networks. On a directed graph one distinguishes between indegree (number of edges incoming to a node) and outdegree (number of edges outgoing from a node), so one can test different types of correlations. The algorithm calculates four correlation functions, namely:

1. $k_{NN}^{in,in}(k)$, where one compares the indegree of a node with the average indegree of their incoming neighbors (i.e. the neighboring nodes adjacent to incoming edges);
2. $k_{NN}^{in,out}(k)$, where one compares the indegree of a node with the average outdegree of their incoming neighbors (i.e. the neighboring nodes adjacent to incoming edges);
3. $k_{NN}^{out,in}(k)$, where one compares the outdegree of a node with the average indegree of their outgoing neighbors (i.e. the neighboring nodes adjacent to outgoing edges);
4. $k_{NN}^{out,out}(k)$, where one compares the outdegree of a node with the average outdegree of their outgoing neighbors (i.e. the neighboring nodes adjacent to outgoing edges);

Each of the above functions is defined as follows. For simplicity of exposition we shall speak in general of degree, neighbors and k_{NN}, keeping in mind the possible combinations. A node is selected and the average degree of all its neighbors is calculated. By repeating the procedure for all nodes of the network one derives a pair (k_{NN}, k) for each node, where k is the degree of the node. By averaging over nodes with equal degree k one derives the function $\bar{k}_{NN}(k)$, which allows to study the correlation. If $\bar{k}_{NN}(k)$ grows with k, the network is assortative; if $\bar{k}_{NN}(k)$ decreases with k, the network is disassortative. A flat curve would indicate the absence of correlation. In the absence of correlations, the function $\bar{k}_{NN}(k)$ would be the same as for a random network with equal degree distribution $P(k)$. In this case it is possible to prove that

1. $\bar{k}_{NN}^{in,in}(k_{in}) = \bar{k}_{NN}^{0,in-in} = \frac{\langle k_{in} k_{out} \rangle}{\langle k_{in} \rangle}$;
2. $\bar{k}_{NN}^{in,out}(k_{in}) = \bar{k}_{NN}^{0,in-out} = \frac{\langle k_{out}^2 \rangle}{\langle k_{out} \rangle}$;
3. $\bar{k}_{NN}^{out,in}(k_{out}) = \bar{k}_{NN}^{0,out-in} = \frac{\langle k_{in}^2 \rangle}{\langle k_{out} \rangle}$;
4. $\bar{k}_{NN}^{out,out}(k_{out}) = \bar{k}_{NN}^{0,out-out} = \frac{\langle k_{in} k_{out} \rangle}{\langle k_{in} \rangle}$.

where $\langle k_{in} k_{out} \rangle$ is the expectation value of the product $k_{in} k_{out}$, $\langle k_{in}^2 \rangle$ is the expectation value of the indegree squared, $\langle k_{out}^2 \rangle$ is the expectation value of the outdegree squared, and $\langle k_{in} \rangle$, $\langle k_{out} \rangle$ the average indegree/outdegree (beware that $\langle k_{in} \rangle = \langle k_{out} \rangle$). For this reason, the functions $\bar{k}_{NN}(k)$ calculated by the algorithm are normalized in that we divide it by the corresponding constants $\bar{k}_{NN}^{0}$.

Links

- Source Code

Applications

The network to analyze must be directed, otherwise there are no special constraints.

The algorithm is used to disclose affinities/diversities between neighboring nodes. Many properties of networks and of processes that take place on networks are affected by the presence of degree-degree correlations.

Implementation Details

The algorithm requires two inputs, the file where the edges of the network are listed and the number of points one wishes to have in the binned correlation function described below. A first read-in of the inputfile will set the values of the number of nodes and edges of the network. In the second read-in the indegrees and outdegrees of all nodes will be calculated and the edges are stored in an array. Then the k_{NN} of all nodes are calculated. The program generates two output files for each correlation function (a total of eight files), corresponding to two different ways of partitioning the interval spanned by the values of degree. In the first output, the k_{NN} is averaged among all nodes with equal degree; the output displays all degree values with their average k_{NN}s (normalized as described in the Description section). The second output gives the binned correlation function, i.e. the interval spanned by the values of degree is divided into bins whose size grows while going to higher values of the variable. The size of each bin is obtained by multiplying by a fixed number the size of the previous bin. The program calculates the average k_{NN} for all nodes whose degree falls within each bin. This technique is particularly suitable to study skewed correlation functions: the fact that the size of the bins grows large for large degree values compensates for the fact that not many nodes have high degree values, so it suppresses the fluctuations that one would observe by using bins of equal size. On a double logarithmic scale the points of $\bar{k}_{NN}(k)$ will appear equally spaced on the x-axis. The program runs in a time O(m), m being the number of edges of the network.

Acknowledgements

The algorithm was implemented and documented by S. Fortunato, integrated by S. Fortunato and W. Huang.

References

Pastor-Satorras, R., Vazquez, A., Vespignani, A. (2001). Dynamical and Correlation Properties of the Internet. Physical Review Letters 87:258701.