Modularity maximization

Our goal is to find a measure that quantifies how many edges lie within groups in our network relative to the number of such edges expected on the basis of chance. A good division of nodes into communities is one that maximizes such a measure. Equivalently, we want a measure that quantifies how many edges lie between groups in our network relative to the expected number of such links. A good division of nodes into communities is one that minimizes such a measure. We will concentrate on the former measure of modularity of a network.

Let us focus on undirected multi-graphs, that is, graphs that allow self-edges (edges involving the same node) and multi-edges (more than one simple edge between two vertices). A measure of modularity of a network is the number of edges that run between vertices of the same community minus the number of such edges we would expect to find if the configuration model is assumed, that is if edges were positioned at random while preserving the vertex degrees. Let us denote c_i the community of vertex i and $\delta(c_i, c_j) = 1$ if $c_i = c_j$ and $\delta(c_i, c_j) = 0$ otherwise. Hence, the number of edges that run between vertices of the same group is:

$$\sum_{(i,j)\in E} \delta(c_i, c_j) = \frac{1}{2} \sum_{i,j} A_{i,j} \delta(c_i, c_j)$$

where E is the set of edges of the graph and $A_{i,j}$ is the actual number of edges between i and j, which is zero or more (notice that each undirected edge is represented by two pairs in the second sum, hence the factor one-half).

The expected number of edges that run between vertices of the same group is:

$$\frac{1}{2} \sum_{i,j} \frac{k_i k_j}{2m} \delta(c_i, c_j)$$

where k_i and k_j are the degrees of i and j, while m is the number of edges of the graph. Notice that $k_i k_j / 2m$ is the expected number of edges between vertices i and j in the configuration model assumption. Indeed, consider a particular edge attached to vertex i. The probability that this edge goes to node j is $k_j/2m$, since the number of edges attached to j is k_j and the total number of edge ends in the network is $2m$ (the sum of all node degrees). Since node i has k_i edges attached to it, the expected number of edges between i and j is $k_i k_j / 2m$.

Hence the difference between the actual and expected number of edges connecting nodes of the same group, expressed as a fraction with respect to the total number of edges m, is called modularity, and given by:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{i,j} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j) = \frac{1}{2m} \sum_{i,j} B_{i,j} \delta(c_i, c_j)$$

where:

$$B_{i,j} = A_{i,j} - \frac{k_i k_j}{2m}$$

and B is called the modularity matrix.

The modularity Q takes positive values if there are more edges between same-group vertices than expected, and negative values if there are less. Our goal is to find the partition of network nodes into communities such that the modularity of the division is maximum. Unfortunately, this is a computationally hard problem. It is believed that the only algorithms capable of always finding the division with maximum modularity take exponentially long to run and hence are useless for all but the smallest of networks. Instead, therefore, we turn to heuristic algorithms, algorithms that attempt to maximize the modularity in an intelligent way that gives reasonably good results in a quick time.

Source: Massimo Franceschet - Università degli Studi "G. d'Annunzio" Chieti – Pescara