Newman's Q

This discussion follows the lines of White and Smyth's article "A Spectral Clustering Approach to Finding Communities in Graphs". It is available online at `www.ics.uci.edu/~scott/siam2005_whitesmyth.pdf`. They cite the original works in Phys Rev E by Newman (2004) and Newman and Girvan (2004).

We begin with a graph $G(\mathcal{V}, \mathcal{E}, W)$ with vertex set $\mathcal{V} = \{1, 2, \ldots, n\}$, edge set $\mathcal{E}$ and a weight matrix W, with $W_{ij} = W_{ji}$ positive if $\mathcal{E}$ contains an edge between nodes i and j and $W_{ij} = 0$ otherwise.

For two subsets $V, V' \subseteq \mathcal{V}$ define $\mathcal{A}(V, V') = \sum_{i \in V} \sum_{j \in V'} W_{ij}$. Let $\mathcal{P}_k$ be a partition of $\mathcal{V}$ into k subsets $V_1, \ldots, V_k$. Then Newman's Q is

$$Q(\mathcal{P}_k) = \sum_{\ell=1}^{k} \left[\frac{\mathcal{A}(V_\ell, V_\ell)}{\mathcal{A}(\mathcal{V}, \mathcal{V})} - \left(\frac{\mathcal{A}(V_\ell, \mathcal{V})}{\mathcal{A}(\mathcal{V}, \mathcal{V})} \right)^2 \right].$$

If we select an edge at random, taking $i \leftrightarrow j$ with probability proportional to W_{ij} then the first part of each term in Q is the probability that an edge begins and ends inside subset V_ℓ of the partition. The second part subtracts the probability that an edge begins inside V_ℓ times the probability that an edge ends inside V_ℓ (those probabilities are equal). Hence Q represents the probability that a random link is within a component of P_k, over and above a measure of the within cluster link probability when links are unrelated to the clusters.

The proposed way to pick the number of clusters is via

$$k^* = \arg\max_k \max_{\mathcal{P}_k} Q(\mathcal{P}_k).$$

Then the resulting clustering is $\mathcal{P}_{k^*}$. In examples the criterion avoids putting all nodes into a single cluster or making each node it's own cluster.

White and Smyth's contribution is a spectral relaxation of the method to increase it's speed. For illustration, they cluster some words by the articles in which they appear, college football teams using a graph with an edge for each game played and NIPS authors via a co-authorship graph.

Statisticians might prefer a ratio like

$$\frac{\sum_{\ell=1}^{k} \frac{\mathcal{A}(V_\ell, V_\ell)}{\mathcal{A}(\mathcal{V}, \mathcal{V})}}{\sum_{\ell=1}^{k} \left(\frac{\mathcal{A}(V_\ell, \mathcal{V})}{\mathcal{A}(\mathcal{V}, \mathcal{V})} \right)^2},$$

to measure the improvement over independent links, but perhaps it's harder to optimize.