Algorithms for Graph Clustering

Algorithm: AgglomerativeClustering(G) - Ravasz Algorithm
Input: Connected graph $G = (V, E)$
Output: Dendrogram whose leaves are the elements of V

1. Assign each node u to its own cluster C_u
2. For all pairs $u, v \in V, u \neq v$, compute their similarity $\text{sim}(u, v)$
3. Repeat until all nodes are in a single cluster:
   (a) Find the pair of clusters C_1, C_2 with the highest similarity $\text{sim}(C_1, C_2)$ (ties are broken arbitrarily)
   (b) Merge clusters C_1, C_2 into a single cluster C'
   (c) Compute similarity between C' and all other clusters
4. Return the corresponding dendrogram

Common choice for $\text{sim}(u, v)$:

$$\text{sim}(u, v) = \frac{|N(u) \cap N(v)| + A_{uv}}{\min\{\deg(u), \deg(v)\} + 1 - A_{uv}}$$

where A is the adjacency matrix of G.

Common choices for $\text{sim}(C_1, C_2)$ are defined different types of linkage clustering:

- **Single linkage clustering:**

$$\text{sim}(C_1, C_2) = \min_{u \in C_1, v \in C_2} \text{sim}(u, v)$$

- **Average linkage clustering:**

$$\text{sim}(C_1, C_2) = \frac{1}{|C_1| \cdot |C_2|} \sum_{u \in C_1, v \in C_2} \text{sim}(u, v)$$

- **Complete linkage clustering:**

$$\text{sim}(C_1, C_2) = \max_{u \in C_1, v \in C_2} \text{sim}(u, v)$$

Algorithm: GNClustering(G) - Girvan-Newman Algorithm
Input: Connected graph $G = (V, E)$
Output: Dendrogram whose leaves are the elements of V

1. Assign all nodes u to a single cluster C
2. Repeat until all nodes are in different clusters:
   (a) For each cluster C:
      i. For each edge $e \in C$, compute $b(e, C)$
   (b) Let $e_{\max}$ be the edge of maximum betweenness, and let $C(e)$ be its cluster
   (c) Remove e from $C(e)$
3. Report the corresponding dendrogram

Modularity based clustering
Definition of Modularity

$$M(S) = \frac{1}{2m} \sum_{u,v \in S} \left(A_{uv} - \frac{\deg(u)\deg(v)}{2m} \right)$$

and

$$M(\mathcal{C}) = \sum_{C \in \mathcal{C}} M(C) = \frac{1}{2m} \sum_{C \in \mathcal{C}} \sum_{u,v \in C} \left(A_{uv} - \frac{\deg(u)\deg(v)}{2m} \right)$$

and

$$\mathrm{sim}(C_i, C_1, C_2) = \frac{|E(C_1, C_2)|}{m} - \frac{\left(\sum_{u \in C_1} \deg(u)\right)\left(\sum_{v \in C_2} \deg(v)\right)}{2m^2}$$

enabling the Algorithm:
GreedyModularityClustering(G)
Input: Connected graph $G = (V, E)$
Output: Clustering of the elements of V

1. Initialize C_1 as the clustering where each node u is assigned to its own cluster C_u; set $i \leftarrow 1$
2. Repeat until all nodes are in a single cluster:
   (a) For each pair of clusters C_1, C_2 such that there exists one edge between C_1 and C_2, compute:
   $$\Delta(C_i, C_1, C_2) = M(C_i \cup C_1 \cup C_2 + (C_1 \cup C_2)) - M(C_i)$$
   (b) Find C_1', C_2' that maximize $\Delta(C_i, C_1', C_2')$
   (c) Update $C_{i+1} \leftarrow C_i \cup C_1' \cup C_2' + (C_1' \cup C_2')$; increment $i \leftarrow i + 1$
3. Return the clustering C^*, across iterations, of maximum modularity:
   $$C^* = \arg \max_{C_i, i=1,2,\ldots} M(C_i)$$