

# CPSC 320 Notes, Clustering Completed

January 30, 2015

(Note: all repeats up to Section 1 below.) We're given a complete, weighted, undirected graph  $G = (V, E)$  represented as an adjacency list, where the weights are all between 0 and 1 and represent similarities—the higher the more similar—and a desired number  $1 \leq k \leq |V|$  of categories.

We define the similarity between two categories  $C_1$  and  $C_2$  to be the maximum similarity between any pair of nodes  $p_1 \in C_1$  and  $p_2 \in C_2$ . We must produce the categorization—partition into  $k$  non-empty sets—that minimizes the maximum similarity between categories.

We'll prove this greedy approach optimal.

1. Sort a list of the edges  $E$  in decreasing order by similarity.
2. Initialize each node as its own category.
3. Initialize the category count to  $|V|$ .
4. While we have more than  $k$  categories:
  - (a) Remove the highest similarity edge  $(u, v)$  from the list.
  - (b) If  $u$  and  $v$  are not in the same category: Merge  $u$ 's and  $v$ 's categories, and reduce the category count by 1.

## 1 Greedy is at least as good as Optimal

We'll start by noting that any solution to this problem partitions the edges into the “intra-category” edges (those that connect nodes within a category) and the “inter-category” edges (those that cross categories).

1. Give a bound on the goodness of a solution in terms of its inter-category edges.
2. Give a bound on the goodness of a solution in terms of its intra-category edges.
3. Consider an optimal solution  $\mathcal{O}$  to an instance of the problem. Prove that its intra-category edges cannot be a proper superset of greedy's intra-category edges. (I.e., it cannot contain all of the intra-category edges of greedy plus at least one more.)

4. Prove that if  $\mathcal{O}$ 's set of intra-category edges is equal to greedy's, then greedy's solution is optimal.
  
5. Prove that if  $\mathcal{O}$ 's set of intra-category edges is neither equal to nor a superset of greedy's, then at least one edge that greedy merged on in step 4(b) is an inter-category edge in  $\mathcal{O}$ .
  
6. Prove that if  $\mathcal{O}$ 's set of intra-category edges is neither equal to nor a superset of greedy's, then greedy's solution is optimal. (Remember: optimal doesn't mean "better than all other solution", just "at least as good as all other solutions".)

## 2 Challenge

1. We can also give an "exchange" argument starting: "Consider a greedy solution  $g$  and an optimal solution  $\mathcal{O}$ . If they're different, then some edge  $(u, v)$  must be inter-category in  $\mathcal{O}$  but intra-category in  $g$ . In that case, we can make  $\mathcal{O}$  more similar to  $g$  without decreasing  $\mathcal{O}$ 's goodness by..."  
 Finish the proof. (Warning: even if you select  $(u, v)$  more carefully, you may not **just** want to move  $u$  into  $v$ 's category or vice versa.)
2. Switching to MSTs:
  - (a) What does the "cut" property tell us if we partition the graph into a "cluster" of: the already connected nodes, and everything else?
  - (b) Use this insight to create an efficient algorithm somewhat similar to Kruskal's algorithm except that it adds **many** edges to the MST-so-far at each step of the outer loop.
  - (c) Analyse the performance of your algorithm, including determining what data structures to use. (Make each data structure as simple as possible while still obtaining the best bound rather than making each as efficient as possible.)