Cluster editing with vertex splittin

Hans Blankensteijn
Problem:

- We want to classify things
- Thinks like proteins! (Proteins are a hot topic atm)
  Cluster editing
  Suppose we have a graph and the nodes in the graph represent proteins. We want the graph to group vertices in cliques if they are similar. Can we do this in only add/remove operatoins.
  Vertex splitting
- This is cluster editing with vertex splitting
- Add an operation with split
  How long does this take?: You can do a reduction 3SAT and it is NP-complete. We are going to look at an algorithm and a greedy heuristinc

Arrighi et al. Workflow

- Calculate critical cliques.
  - A critical clique is a Group of vertices where all vertices share the same neighborhoods and is maximal under this property
  - Each vertex belongs to only 1 CC
  - amount of CC<4k
- Color the vertices
  - Let $\chi = (S_1, \ldots, S_l)$ be the set of cliques in G'
  - Each operation can complet ate most 2 cliques, so $l \leq 2k$ (if $l \geq 2k$, we cannot solve in k operations)
  - We are going to color our vertices wit $l + 1$ colors
  - Color 0 will be a split-vertex
  - Rest of the vertices will be colored at random
  - There are $(l + 1)^{4k} \in O((2k + 1)^{4k})$ colorings
  - Try them all
- Guessing splits
  - We have vertices of colour 0. In what clique will the land?
  - Approach: guess that 0 colored vertex i is in clique j
  - $(kl + 1)^k = (2k^2 + 1)^k$ guesses
- Performing operations
  - If color is equal add edge (u,v)
  - If color is unequal remove edge (u,v

$O(n+m)$ time for making CC's

Greedy heuristic

- Greedy: do operations of lowest cost first
- cost add(uv)=#operations to connect N(u) and N(v)
- cost delete(uv) =#operations to disconnect N(u) and N(v)
- set uv to permanent or forbidden
- also: predict permanent and forbidden edges beforehand (use threshold)
- Make use of conflict triples.
- Set uv to permanent if they share neighbors
- Set uv to forbidden if the don't share neighbors
- If uv and vw are permanent, but uw is forbidden: split v
- Use boolean to regulate the amount of splitting
- What to do with neighbors?

The algorithms works nicely. They compared it to k-means clustering and they discovered it gives a better solution, but is a bit slower.

Conclusion

- Clustering is NP-hard
- Cluster editing is a graphical way of looking at clustering data points
- when devising graph algorithms make use of graph properties
- Exact algorithms give insights to the intricacies of a problem
- Sometimes heuristics can be good enough