Ensemble Clustering using Semidefinite Programming W2

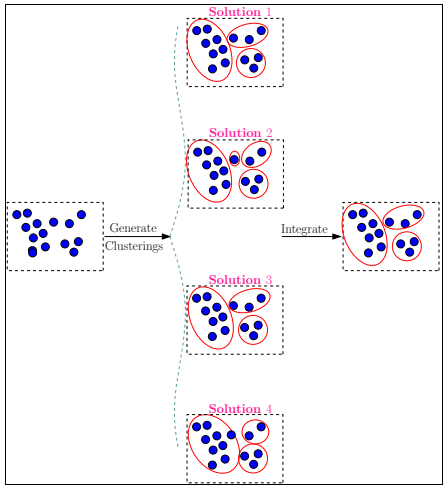
Vikas Singh[1], Lopamudra Mukherjee[2], Jiming Peng[3], Jinhui Xu[2]

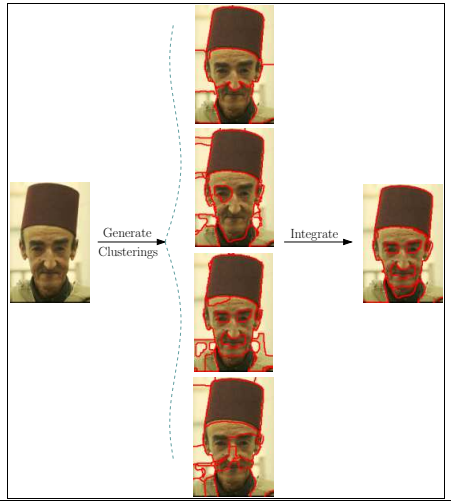
[1]Biostatistics and Medical Informatics
University of Wisconsin – Madison
vsingh@biostat.wisc.edu

[2]Computer Science & Engineering
State University of New York at Buffalo
{lm37, jinhui}@cse.buffalo.edu

[3]Industrial and Enterprise Systems Eng.
University of Illinois Urbana Champaign
pengj@uiuc.edu

The Problem: How to best combine the **ensemble of multiple clustering solutions** in a maximum consent sense?





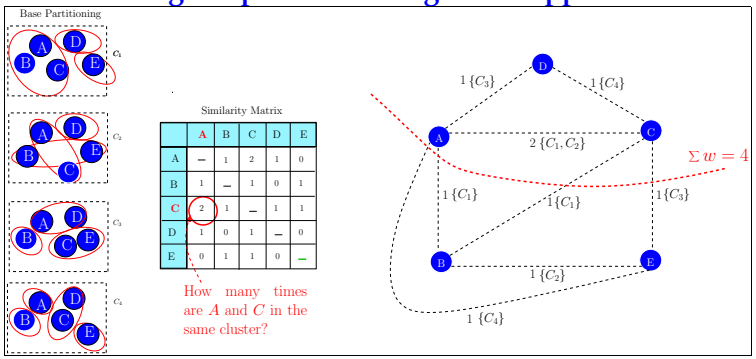
Goal

- Let $P_1, P_2, \ldots, P_m$ be m partitions of the data.
- Each partition P_i is produced by a different algorithm, C_i in an ensemble.
- Goal is to derive a partition P^* for the ensemble.

Motivations

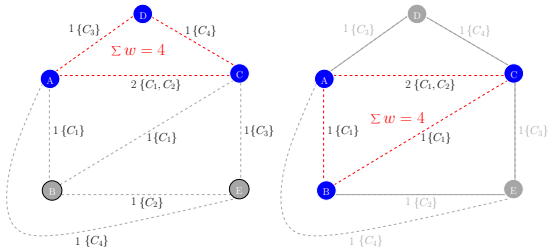
- No single clustering algorithm is perfect.
- In most real life applications, no ground truth is known.
- An ensemble will most likely be superior to individual solutions.

Existing Graph Partitioning based approaches



Compute togetherness frequency/votes/weight of items, but considered pairwise

Do pairwise voting strategies work?



(A, C, D) and (A, B, C) are same weight-wise, **but**

Which algorithms put (A, C, D) together?
Answer: None
Which algorithms put (A, B, C) together?
Answer: At least one (C_1)
Is this information useful?
Answer: **Yes**, as cohesiveness measures in clusters of size ≥ 2

Our contributions

- **A new string encoding based strategy captures ← this cohesiveness/agreement better.**
- **A new IP model to optimize similarity.**
- **Model can be convexified to a precise SDP.**
- **Show applications to novel domains (such as segmentation) in addition to regular classification tasks.**
- **Empirically, also optimizes Normalized Mutual Information well.**

So, we'll see you at the poster then!

Advances in Neural Information Processing Systems, 2007.