

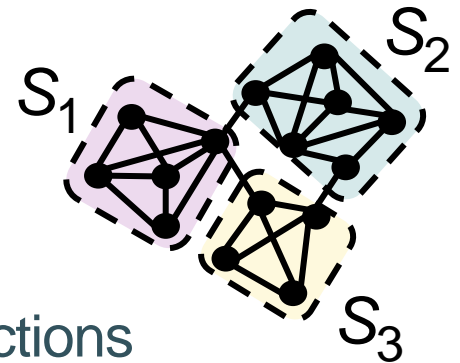
# Minimum Average Cost Clustering

\* Kiyohito Nagano (University of Tokyo)  
Yoshinobu Kawahara (Osaka University)  
Satoru Iwata (Kyoto University)

For clustering problems with submodular objective functions, we introduce the minimum average cost criterion

The proposed algorithm

- does not require # of clusters in advance
- computes an optimal # of clusters and an optimal partition in polynomial time
- uses the theory of intersecting submodular functions



## Keywords

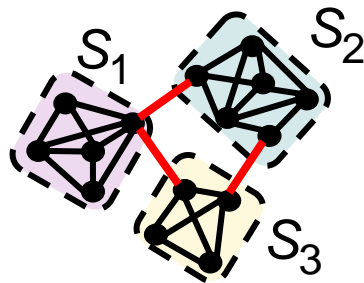
clustering, [submodular functions](#), combinatorial optimization

- $V = \{1, \dots, n\}$  is a finite set of data points
- A function  $f$  defined on  $2^V = \{S: S \subseteq V\}$  is **submodular** if
 
$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T), \quad \forall S, T \subseteq V$$
  - A generalization of **cut functions**, **entropy functions**, etc.

## Clustering problem with submodular objective function

[Narasimhan-Jojic-Bilmes, NIPS 2005]

Given set  $V$ , integer  $k (\leq n)$ , and submodular function  $f$ ,  
 find a  $k$ -partition of  $V$ ,  $\{S_1, \dots, S_k\}$  that minimizes  $\sum_{i=1}^k f(S_i)$



optimal 3-clustering

optimal  $k$ -clustering

In the case of a network,  $\sum_i f(S_i) = 2 \times \#(\text{red edge})$   
 ( $V$  is a set of nodes, and  $f$  is a cut function)

## ■ Optimal $k$ -clustering problem [Narasimhan *et al.*, NIPS 2005]

$$\begin{aligned} \min \quad & \sum_{i=1}^k f(S_i) \\ \text{s. t.} \quad & \{S_1, \dots, S_k\} \text{ is a } k\text{-partition of } V \end{aligned}$$

○  $k$  (# of clusters) should be computed via some method

☹ NP-hard

## ■ Minimum Average Cost (MAC) clustering [This work]

$$\begin{aligned} \min \quad & \sum_{S \in \mathcal{P}} f(S) / (|\mathcal{P}| - \beta) \\ \text{s. t.} \quad & \mathcal{P} \text{ is a partition of } V \\ & |\mathcal{P}| > \beta \end{aligned}$$

averaged objective function

If  $\beta = 1$ , then the objective function is a natural average cost of  $\mathcal{P}$

where  $0 \leq \beta < n$

$\beta$ -MAC clustering

○  $k = |\mathcal{P}|$  and a partition  $\mathcal{P}$  are determined at the same time

😊 Solvable in poly time

😊 competitive with other methods

If  $\beta$  is small,  $\mathcal{P}$  is coarse. If  $\beta$  is big,  $\mathcal{P}$  is fine.

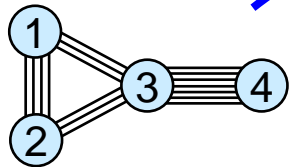
## Theory of intersecting submodular functions

⇒ **Theorem** [This work]. There is an algorithm that computes all the  $\beta$ -MAC clusterings in **polynomial time** in total

**Observation.** Suppose that a partition  $\mathcal{P}$  is a  $\beta$ -MAC clustering for some  $\beta$ , and let  $k = |\mathcal{P}|$ . Then,  $\mathcal{P}$  is a  $k$ -optimal clustering

⇒ The information about MAC clusterings gives a portion of the information about optimal  $k$ -clusterings

(Remember that an optimal  $k$ -clustering problem is NP-hard)

**Example**

Compute  
all  $\beta$ -MAC  
clusterings

$$0 \leq \beta < 1$$

$$\{\{1, 2, 3, 4\}\}$$

opt 1-clustering

$$1 \leq \beta < 11/7$$

$$\{\{1\}, \{2\}, \{3, 4\}\}$$

opt 3-clustering

$$11/7 \leq \beta < 4$$

$$\{\{1\}, \{2\}, \{3\}, \{4\}\}$$

opt 4-clustering

In this case, our algorithm computes optimal  $k$ -clusterings for  $k = 1, 3, \& 4$

For more information,  
please visit Poster T26