Integer and Combinatorial Optimization: Clustering Problems

John E. Mitchell

Department of Mathematical Sciences
RPI, Troy, NY 12180 USA

February 2019
We have $n$ objects, each with a number of attributes.

We wish to group similar objects into clusters.

There is no limit on the number of clusters, or on the size of each cluster.

We have a measure $c_{ij}$ of the difference between two objects $i$ and $j$; the larger this measure, the less similar the objects.

This measure can take **positive or negative** values.
Variables and dimension

We model this by introducing variables

\[ x_{ij} = \begin{cases} 
1 & \text{if } i \text{ and } j \text{ in same cluster} \\
0 & \text{if } i \text{ and } j \text{ in different clusters} 
\end{cases} \quad \text{for } 1 \leq i < j \leq n \]

Let \( S \subseteq \mathbb{B}^{\frac{1}{2}}n(n-1) \) be the set of feasible solutions. We have the following results regarding \( \text{conv}(S) \):

**Proposition**

*The set \( S \) is full-dimensional.

One way to prove this is to note that the origin and all the unit vectors are in \( S \).
Nonnegativity

**Proposition**

*The lower bound constraints* $x_{ij} \geq 0$ *define facets of* $\text{conv}(S)$.\*
Triangle inequalities

**Proposition**

Let $1 \leq i < j < k \leq n$. The triangle inequalities

\[
\begin{align*}
    x_{ij} + x_{ik} - x_{jk} & \leq 1 \\
    x_{ij} - x_{ik} + x_{jk} & \leq 1 \\
    -x_{ij} + x_{ik} + x_{jk} & \leq 1
\end{align*}
\]

*define facets of $\text{conv}(S)$.*

These inequalities enforce consistency.

For example, the first one says that if $i$ and $j$ are in the same cluster and also $i$ and $k$ are in the same cluster then $j$ and $k$ must be in the same cluster. The only binary solution violating this constraint is $x_{ij} = x_{ik} = 1, x_{jk} = 0$. 
An integer program

Proposition

Any binary vector satisfying all the triangle inequalities is the incidence vector of a clustering.

Thus, finding the best binary vector satisfying the triangle inequalities will solve the clustering problem.
The upper bound constraints $x_{ij} \leq 1$ do not define facets of $\text{conv}(S)$. In particular, if $x_{ij} = 1$ then we must also have $x_{ij} + x_{ik} - x_{jk} = 1$ and $x_{ij} - x_{ik} + x_{jk} = 1$ for each other $k$. Must have $x_{ik} = x_{jk}$.
2-partition inequalities

The following proposition generalizes the lower bound and triangle inequalities.

Proposition

\[(2\text{-partition inequalities}) \text{ Let } U \text{ and } W \text{ be disjoint collections of objects with } |U| > |W|. \text{ The following inequality defines a facet of } \text{conv}(S):\]

\[
\sum_{i \in U, j \in W} x_{ij} - \sum_{i \in U, j \in U} x_{ij} - \sum_{i \in W, j \in W} x_{ij} \leq |W|.
\]

This gives the lower bound constraints when \(|U| = 2, |W| = 0\). It gives the triangle constraints when \(|U| = 2, |W| = 1\).
\[ \sum_{e = (i,j): i \in U, j \in W} x_{ij} - \sum_{e = (i,j): i \in W, j \in U} x_{ij} \leq |W| \]

\[ x_{uv} + x_{vt} - x_{uv} = 1 \quad \text{triangle inequality} \]

\[ \sum_{u \in U \cup \emptyset} w = 4 \quad -x_{uv} \leq 0 \]
The objective function coefficients

Note that if all the $c_{ij}$ are nonnegative then the optimal solution is to place each object in its own cluster, so all $x_{ij} = 0$.

Thus, our measure $c_{ij}$ cannot simply be the distance between two objects, but must allow negative values if we are to have an interesting problem.

For more details, see Grötschel and Wakabayashi [3, 4].
Given a graph $G = (V, E)$ with $n = |V| = 2q$ for some integer $q$, we partition $V$ into two sets of size $q$. We define the variables

$$x_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ in same partition} \\ 0 & \text{if } i \text{ and } j \text{ in different partitions} \end{cases} \text{ for } 1 \leq i < j \leq n$$

Let $S$ be the set of feasible incidence vectors of equipartitions.

Costs $c_{ij}$ on edges.

Objective: either $\min \sum_{e \in E} c_e x_e$

or $\max \sum_{e \in E} c_e x_e$
Polyhedral results

We have the following results:

**Proposition**

The dimension of $\text{conv}(S)$ is $\frac{1}{2}n(n - 3)$.

If $C$ is a cycle with $q + 1$ vertices then the inequality $x(E(C)) \leq q - 1$ is facet defining.

If $U \subseteq V$ with $|U| \geq 3$ and odd, the clique inequality $x(E(U)) \geq \left(\frac{1}{2}|U|\right)^2$ is facet-defining.

Other inequalities are known (Conforti et al. [1, 2]).
Clustering with lower bound

Now consider a clustering problem where we require each cluster to contain at least $q$ elements, for some positive integer $q$.

For example, this problem arises in the following settings:

- **allocating teams to divisions in a sports league.** In this case, often require each division to have the same cardinality.
- **microaggregation in the release of data**: in order to preserve privacy, clusters with tiny sizes must be avoided.
Clustering with lower bound

Now consider a clustering problem where we require each cluster to contain at least $q$ elements, for some positive integer $q$.

For example, this problem arises in the following settings:

- **allocating teams to divisions in a sports league.** In this case, often require each division to have the same cardinality.
- **microaggregation in the release of data:** in order to preserve privacy, clusters with tiny sizes must be avoided.
Polyhedral theory

Let $S \subseteq \mathbb{B}^{\frac{1}{2}n(n-1)}$ be the set of incidence vectors of clusterings where each cluster contains at least $q$ elements. We have the following results regarding conv($S$):

**Proposition**

If $q < n/2$ then $\dim(\text{conv}(S)) = \frac{1}{2}n(n - 1)$, so $S$ is full-dimensional.

**Proposition**

The nonnegativity constraints and the triangle constraints of Proposition 3 define facets of conv($S$), provided $q < n/3$. The 2-partition inequalities of Proposition 5 define facets of conv($S$) provided $(|W| + 2)q < n$.

Other families of valid inequalities are also known [5, 6].
Polyhedral theory

Let $S \subseteq \mathbb{B}^{\frac{1}{2}n(n-1)}$ be the set of incidence vectors of clusterings where each cluster contains at least $q$ elements. We have the following results regarding $\text{conv}(S)$:

**Proposition**

If $q < n/2$ then $\dim(\text{conv}(S)) = \frac{1}{2} n(n - 1)$, so $S$ is full-dimensional.

**Proposition**

The nonnegativity constraints and the triangle constraints of Proposition 3 define facets of $\text{conv}(S)$, provided $q < n/3$. The 2-partition inequalities of Proposition 5 define facets of $\text{conv}(S)$ provided $(|W| + 2)q < n$.

Other families of valid inequalities are also known [5, 6].
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.

M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope II: Valid inequalities and facets.

M. Grötschel and Y. Wakabayashi.
A cutting plane algorithm for a clustering problem.

M. Grötschel and Y. Wakabayashi.
Facets of the clique partitioning polytope.

X. Ji and J. E. Mitchell.
Branch-and-price-and-cut on the clique partitioning problem with minimum clique size requirement.

J. E. Mitchell.
Realignment in the national football league: Did they get it right?
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.  

M. Conforti, M. R. Rao, and A. Sassano.  
The equipartition polytope II: Valid inequalities and facets.  

M. Grötschel and Y. Wakabayashi.  
A cutting plane algorithm for a clustering problem.  

M. Grötschel and Y. Wakabayashi.  
Facets of the clique partitioning polytope.  

X. Ji and J. E. Mitchell.  
Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement.  

J. E. Mitchell.  
Realignement in the national football league: Did they get it right?  
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.

M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope II: Valid inequalities and facets.

M. Grötschel and Y. Wakabayashi.
A cutting plane algorithm for a clustering problem.

M. Grötschel and Y. Wakabayashi.
Facets of the clique partitioning polytope.

X. Ji and J. E. Mitchell.
Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement.

J. E. Mitchell.
Realignment in the national football league: Did they get it right?
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.  

M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope II: Valid inequalities and facets.  

M. Grötschel and Y. Wakabayashi.
A cutting plane algorithm for a clustering problem.  

M. Grötschel and Y. Wakabayashi.
Facets of the clique partitioning polytope.  

X. Ji and J. E. Mitchell.
Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement.  

J. E. Mitchell.
Realignment in the national football league: Did they get it right?  
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.

M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope II: Valid inequalities and facets.

M. Grötschel and Y. Wakabayashi.
A cutting plane algorithm for a clustering problem.

M. Grötschel and Y. Wakabayashi.
Facets of the clique partitioning polytope.

X. Ji and J. E. Mitchell.
Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement.

J. E. Mitchell.
Realignment in the national football league: Did they get it right?
M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope I: Formulations, dimension and basic facets.

M. Conforti, M. R. Rao, and A. Sassano.
The equipartition polytope II: Valid inequalities and facets.

M. Grötschel and Y. Wakabayashi.
A cutting plane algorithm for a clustering problem.

M. Grötschel and Y. Wakabayashi.
Facets of the clique partitioning polytope.

X. Ji and J. E. Mitchell.
Branch-and-price-and-cut on the clique partition problem with minimum clique size requirement.

J. E. Mitchell.
Realignment in the national football league: Did they get it right?