Socially Fair Clustering

Instructor: Yury Makarychev, TTIC
Fairness in Facility Location

Choose locations for stores/hospitals/fire stations/etc so as to minimize the average distance from people to these facilities.

+ fair for minority groups
Fair clustering

Given:
• a set of points $X$ and a distance function $d$ on $X$.
• a list of groups $G_1, \ldots, G_\ell \subset X$

Centers and clustering:
A set of centers $\{c_1, \ldots, c_k\}$ defines the Voronoi clustering: cluster $C_i$ consists of the points that are closer to $c_i$ than to other centers

Cost function:
Let
\[
\text{cost}(j, C) = \frac{1}{|G_j|} \sum_{u \in G_j} d(u, C)^p.
\]
\[
\text{cost}(C) = \max_{1 \leq j \leq \ell} \text{cost}(j, C).
\]
Known Results for $k$-medians and $k$-means

$k$-medians:

$\frac{2}{3}$

Charikar, Guha, Tardos, Shmoys ‘02

2.675

Byrka, Pensyl, Rybicki, Srinivasan, Trinh ‘14

$k$-means:

6.357

Ahmadian, Norouzi-Fard, Svensson, Ward ‘17
Known Results

In the context of socially fair clustering, the problem was introduced by

- Abbasi, Bhaskara, and Venkatasubramanian (2021) for \( p = 1, 2 \)
- Ghadiri, Samadi, and Vempala (2021) for \( p = 2 \)

They gave

- an \( O(\ell) \) approximation algorithm
- a matching integrality gap of \( \Omega(\ell) \)
- a bicriteria approximation algorithm

Anthony, Goyal, Gupta, and Nagarajan (2010) studied the problem in the context of “robust clustering” and gave an \( O(\log n + \log \ell) \) approximation algorithm for \( p = 1 \).
Known Results

Bhattacharya, Chalermsook, Mehlhorn, and Neumann (2014): The problem doesn’t admit a better than \( O\left(\frac{\log \ell}{\log \log \ell}\right) \) approximation unless \( NP \subset \bigcap DTIME(2^{n^\delta}) \).

M, Vakilian (2021): There is an \( O\left(\frac{\log \ell}{\log \log \ell}\right) \) approximation algorithm for every \( p \) (the constant in \( O(\cdot) \) depends on \( p \)).
**Our Setting**

Original setting:

\[
\text{cost}(j, C) = \frac{1}{|G_j|} \sum_{u \in G_j} d(u, C)^p
\]

More general setting:

\[
\text{cost}(j, C) = \sum_{u \in X} w_j(u) \cdot d(u, C)^p
\]

In particular, we may let

\[
w_j(u) = \frac{1}{|G_j|} \quad \text{if} \quad u \in G_j \quad \text{and} \quad ... = 0, \quad \text{otherwise}
\]
Basic LP Relaxation

LP variables

\[ x_{uv} \] is the indicator variable of the event that \( u \) is assigned to center \( v \)

\[ y_v \] is the indicator variable of the event that \( v \) is a center
\[ y_a = y_b = y_c = y_d = 1 \]
\[ y_u = y_v = \cdots = 0 \]

\[ x_{ua} = x_{vb} = \cdots = 1 \]
\[ x_{ub} = x_{uc} = x_{ud} = 0 \]
minimize $z$

s.t.

$z \geq \sum_{uv} w_j(u) d(u, v) \cdot x_{uv}$ for all $j = 1, \ldots, \ell$

$\sum_v x_{uv} = 1$ every $u$ is assigned to some center

$\sum_v y_v \leq k$ there are at most $k$ centers

$x_{uv} \leq y_v$ $u$ is assigned to center $v$, only if $v$ is a center

$x_{uv}, y_v \geq 0$