

Approximation Algorithms For Fair Range Clustering

Sèdjro Hotegni

African Institute for Mathematical
Sciences—Rwanda



Sepideh Mahabadi

Microsoft Research—Redmond



Ali Vakilian

Toyota Technological Institute at
Chicago (TTIC)

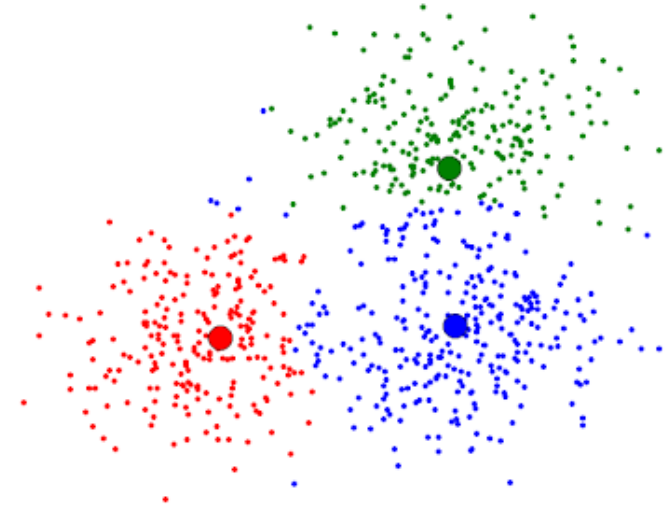


Clustering: A FUNDAMENTAL COMPUTATIONAL TASK

High-Level Goal. Grouping **a set of objects** so that ones in the same group are more **“similar to”** each other than to those in other groups (i.e., determined by a **distance function**).

Centroid-Based Clustering Formulation: pick k centers C to minimize a given cost function

- **k -center:** *min* maximum distance of any point to centers
 ℓ_∞ -objective
- **k -median:** *min* sum of distances of points to centers
 ℓ_1 -objective
- **k -means:** *min* sum of squared distances of points to centers
 ℓ_2 -objective



More generally, for $p \in [1, \infty)$, minimize
 ℓ_p -objective = $(\sum_{v \in P} d(v, C)^p)^{1/p}$

Centroid-Based Clustering: FAIRNESS

DATA SUMMARIZATION

Centers are picked as a summary of the whole dataset.

- Text Summarization (e.g., in legal cases)
- Search Results

- Percentage of U.S. CEOs who are women $\approx 30\%$

Need to enforce the fairness as a requirement in the center selection process:

- New optimization (i.e., clustering) problem

The screenshot shows a Google search for "CEO United States". The search bar is at the top, and below it are navigation tabs for "All", "Images", "News", "Shopping", "Maps", and "More". Below the navigation tabs are several search filters: "pay", "humane society", "steel", "average age", and "compensation". The search results are displayed in a grid format. The first row includes three results: "Chief executive officer - Wikipedia", "CEO pay: Heads of Microsoft, Intel and ...", and "The 15 States With the Most CEOs". The second row includes three results: "The 20 Richest CEOs in the United States", "Should Pay Attention To During COVID-19", and "U.S. Steel CEO David Burritt talks ...". The third row includes three results: "Coal miner Peabody's CEO Kellow to step...", "Growth of CEO pay in America" (a line graph showing the ratio of average CEO compensation to average production worker compensation from 1996 to 2008), and "Wilson Center Names Ambassador Mark...". At the bottom of the screenshot, there is a large white text overlay that reads "SEARCH: 'CEO United States'" over a dark background.

Centroid-Based Clustering: FAIRNESS

A notion of fair clustering proposed by [Kleindessner, Awasthi, Morgenstern, ICML'19]:

Input: k_i where $k := \sum_{i \in [\ell]} k_i$,

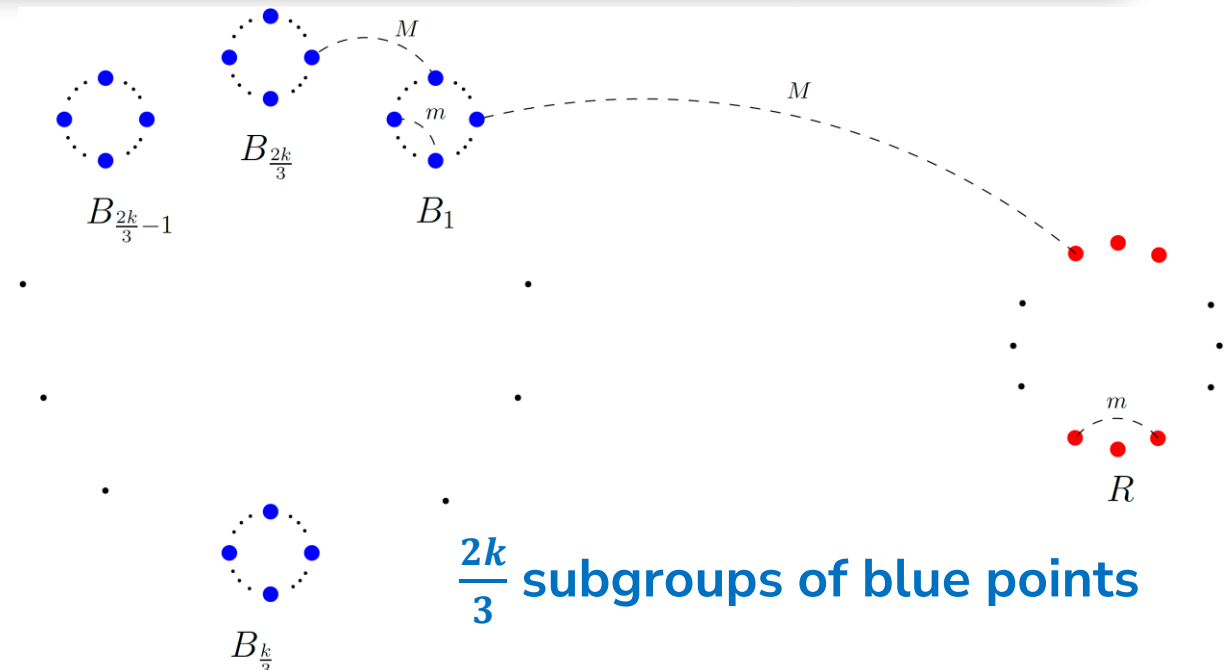
Goal: pick centers C with minimum k -center cost s.t. $\forall i \in [\ell], |C \cap P_i| = k_i$

- Plausible fix for unfairness issue
- Significant loss in the clustering quality

- k -center cost with no constraint: m
- k -center cost with $k_{\text{blue}} = k_{\text{red}}$: $M \gg m$

RELAXED NOTION

- k -center cost with $k_{\text{red}}, k_{\text{blue}} \in [\frac{k}{3}, \frac{2k}{3}]$: m



Clustering: FAIR RANGE FORMULATION

FAIR RANGE CLUSTERING: Given a set of n points in a metric space (P, d) :

- Each point belongs to one of given ℓ different groups ($P = P_1 \uplus P_2 \uplus \dots \uplus P_\ell$)
 - Set of ℓ intervals $[\alpha_1, \beta_1], \dots, [\alpha_\ell, \beta_\ell]$
 - Pick k centers C with minimum clustering cost s.t. $\forall i \in [\ell], \alpha_i \leq |C \cap P_i| \leq \beta_i$
-
- Generalizes the notion of [\[Kleindessner et al., ICML'19\]](#): $\alpha_i = \beta_i = k_i$
 - Each group is at least minimally represented in the center set: $|C \cap P_i| \geq \alpha_i$
 - No group dominates the center set: $|C \cap P_i| \leq \beta_i$

Fair Range Clustering: BACKGROUND

Fair range clustering proposed by [Nguyen, Nguyen, Jones'22]:

- $O(1)$ -approximation for k -center objective
- Not for other standard objectives such as k -median and k -means and more generally ℓ_p -cost: $\min_C (\sum d(v, C)^p)^{1/p}$

Our Contribution:

$O(1)$ -approximation for fair range clustering with ℓ_p -cost where $p \in [1, \infty)$

- LP-based approach
- The LPs are solvable in time $(nk)^{1.5}$

High-Level Overview of Our Algorithm

RELAXATION

Step 1. Find an optimal fractional solution of a natural LP relaxation

REDUCTION TO A SPARSE INSTANCE

Step 2. Reduce the input instance to a *sparse instance*

SPARSITY. The new instance is supported on $O(k)$ points

STRUCTURED. It admits a “good” *structured fractional solution* (e.g., *almost integral*)

ROUNDING THE SPARSE SOLUTION

Step 3. Round the fractional solution of the *sparse instance*

Write new LP-relaxation for the sparse structured instance

Find a half-integral optimal solution of the new LP

Round the half-integral solution via an application of Max-Flow