Approximation Algorihtms For Fair Range Clustering

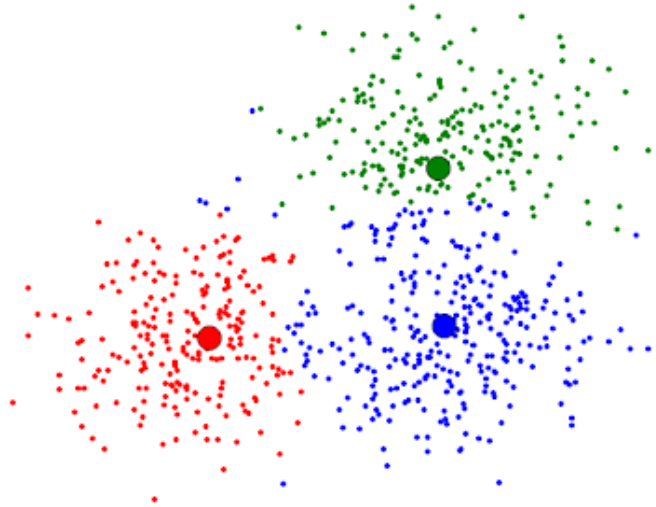
Sèdjro Hotegni
African Institute for Mathematical Sciences—Rwanda

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

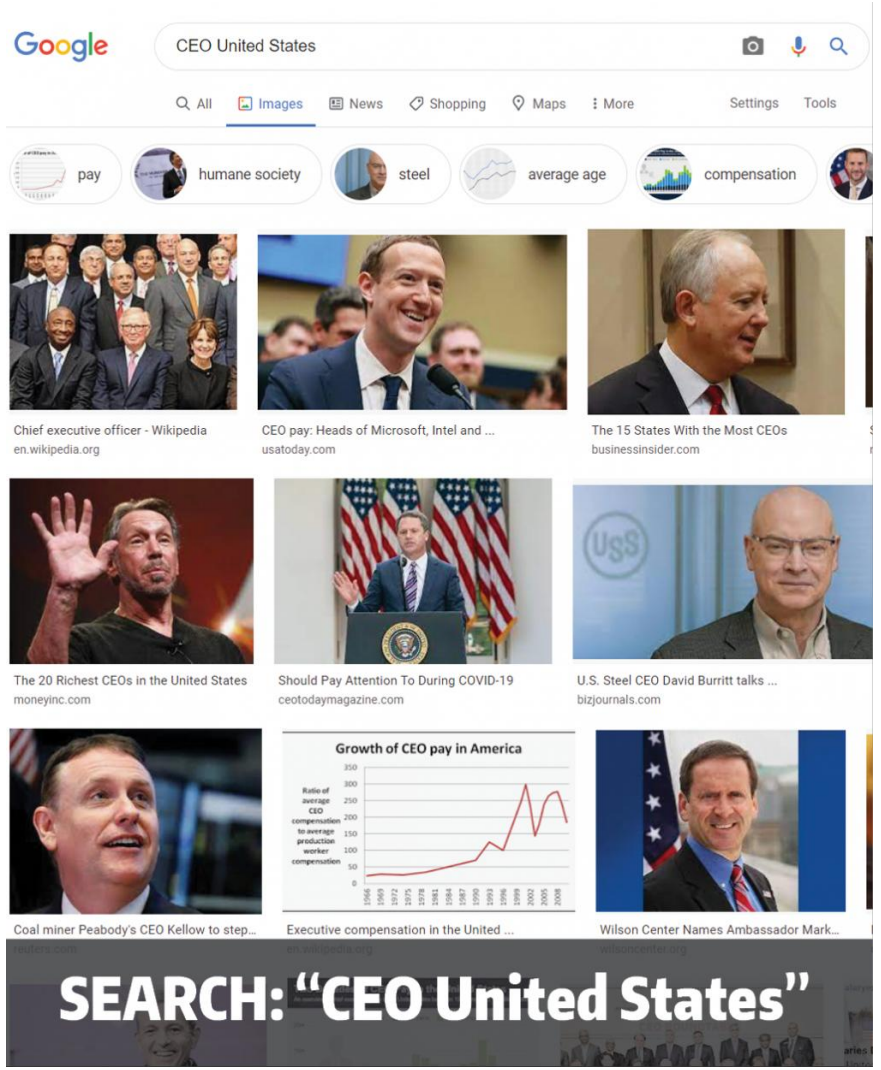
Centers are picked as a summary of the whole dataset.

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

- Percentage of U.S. CEOs who are women ≈ 30%

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

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



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. $\forall i \in [\ell], |C \cap P_i| = k_i$

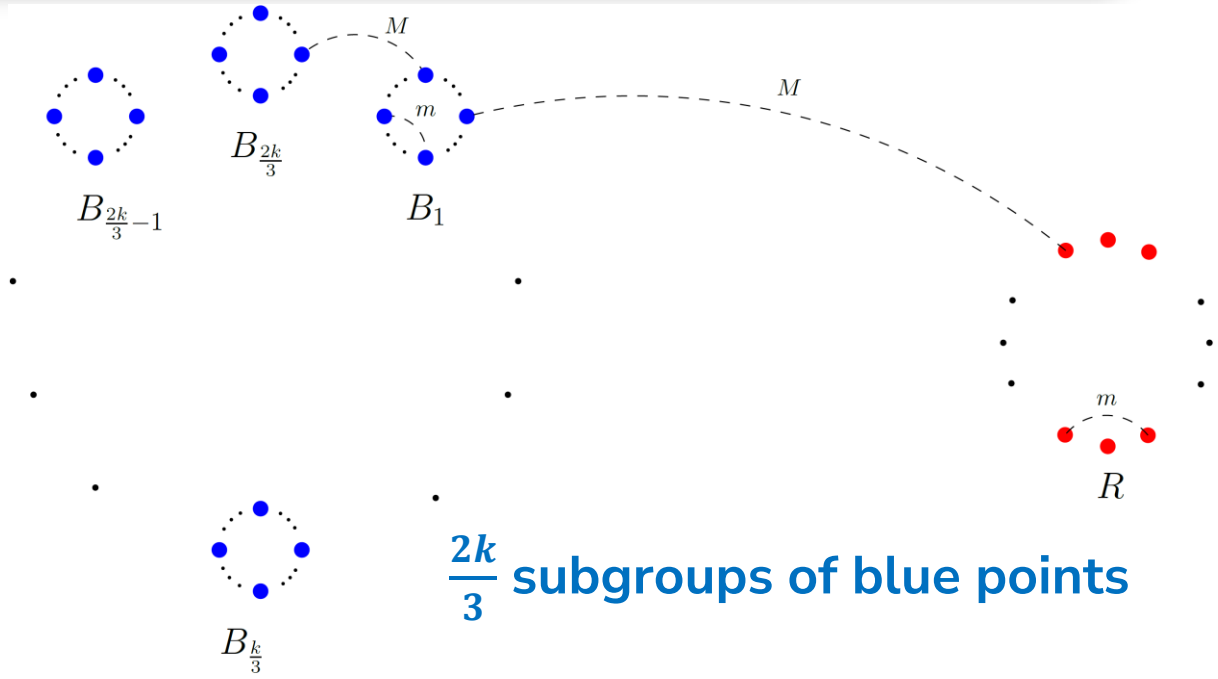
- Plausible fix for unfairness issue
- Significant loss in the clustering quality

❑ 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



$\frac{2k}{3}$ subgroups of blue points

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 ($P = P_1 \uplus P_2 \uplus \cdots \uplus P_\ell$)
- Set of ℓ intervals $[\alpha_1, \beta_1], \cdots, [\alpha_\ell, \beta_\ell]$
- 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