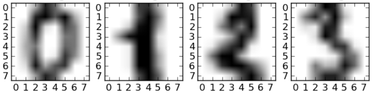
Scalable and Improved Algorithms for Individually Fair Clustering

Hossein Bateni Vincent Cohen-Addad Alessandro Epasto Silvio Lattanzi

Google Research

Metric Clustering

Goal: Partition data according to *similarity*.



Underlying data: Points in $\mathbb{R}^2$.

Other examples of low-dimensional inputs: Image segmentation, facility location, etc.

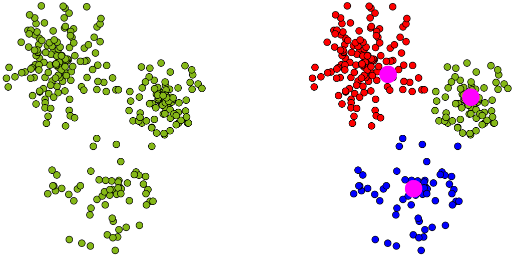
Metric Clustering Objectives

k-Clustering

Input: *data points* A in a metric space
Output: set C of k *centers* that minimizes

$$\sum_{a \in A} \min_{c \in C} \mathrm{dist}(a, c)^p.$$

k-median is when $p = 1$, k-means is when $p = 2$.



In Practice: k-means objective more popular than k-median

Fair Clustering

Individual elements are of different *types*.

Fair Clustering

Goal: Center quality is the same for all types.
No type has a much larger distance to the centers.

Fair k-Median

For each point p, closest center must be at distance at most $\delta(p)$.

Bicriteria Approximation

α, β-approximation $\iff$ α-approximation of the k-median cost and constraints are satisfied up to a factor β.

Fair k-Median

Credit	k-center or fairness guarantee	k-median guarantee	Runtime
Alamdari & Shmoys	4^*	8	polynomial
Humayun et al.	-	-	$\Omega((n^2k)^{2.37})$
Mahabad & Vakilian	7	84	$\tilde{O}((kn)^5/\epsilon)$
Chakrabarty & Swamy	8	8	$\tilde{O}(kn^4)$
Vakilian & Yalçiner	3	$8+\epsilon$	$\Omega(n^4)$
This work ($\gamma \geq 6$)	$\gamma + 1$	$3 + O(\varepsilon)$	$n^{O(1/\epsilon)}$
This work ($6 > \gamma > 4$)	$\gamma + 1$	$\frac{3\gamma-2}{2\gamma-8} + O(\varepsilon)$	$n^{O(1/\epsilon)}$
This work	6	$O(1)$	$\tilde{O}(nk^2)$

Our Results

Theorem

Let $\gamma > 4$ and $\varepsilon > 0$. Assuming the problem is feasible (i.e., there exists an individually fair solution), there is a polynomial-time algorithm for individually fair k-median with bicriteria guarantee $(\alpha_\gamma, \gamma+1)$, where $\alpha_\gamma = 3+O(\varepsilon)$ for $\gamma \geq 6$ and $\alpha_\gamma = \frac{2+\frac{4}{\gamma-2}}{2-\frac{4}{\gamma-2}} + O(\varepsilon)$ for $6 > \gamma > 4$.

Anchored local Search algorithm:

$S_0 \leftarrow$ Gonzales Algorithm finds solution that satisfies the cstrt up to a factor γ.
$S \leftarrow S_0$
While there exists a solution S' such that
 $\text{cost}(S') < (1 - 1/n)\text{cost}(S)$; and
 $|S \setminus S'| + |S' \setminus S| \leq 2/\epsilon$; and
 and for each $p \in S_0$, $|S' \cap B(p, \delta(p))| \neq \emptyset$:
$S \leftarrow S'$
output S

Our Results

Theorem

Let $\gamma > 4$ and $\varepsilon > 0$. Assuming the problem is feasible (i.e., there exists an individually fair solution), there is a polynomial-time algorithm for individually fair k-median with bicriteria guarantee $(\alpha_\gamma, \gamma+1)$, where $\alpha_\gamma = 3+O(\varepsilon)$ for $\gamma \geq 6$ and $\alpha_\gamma = \frac{2+\frac{4}{\gamma-2}}{2-\frac{4}{\gamma-2}} + O(\varepsilon)$ for $6 > \gamma > 4$.

Anchored local Search algorithm:

$S_0 \leftarrow$ Gonzales Algorithm finds solution that satisfies the cstrt up to a factor γ.

$S \leftarrow S_0$

While there exists a solution S' such that

$\text{cost}(S') < (1 - 1/n)\text{cost}(S)$; and

$|S \setminus S'| + |S' \setminus S| \leq 2/\epsilon$; and

and for each $p \in S_0$, $|S' \cap B(p, \delta(p))| \neq \emptyset$:

$S \leftarrow S'$

output S

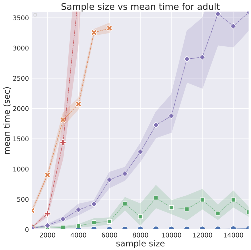
Our Results: A Faster Algorithm

Theorem

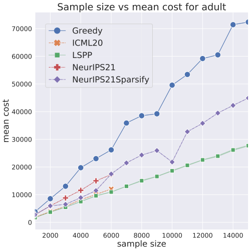
There is an $\tilde{O}(nk^2)$-time algorithm for individually fair k-means with a 6-approximation for radii and an $O(1)$-approximation on costs.

Idea: Anchoring as above and $k\texttt{-means++}$ sampling of centers to replace.

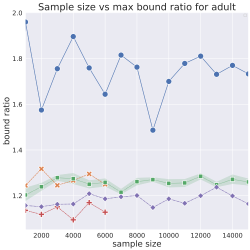
Our Experimental Results



(a) Time (secs)



(b) Cost



(c) Bound ratio

Mean completion time, cost, and bound ratio for the algorithms on adult dataset subsampled to different sizes, $k = 10$. The shades represent the 95% confidence interval (notice that some algorithms are deterministic). Runs that did not complete in 1 hour are not reported.

Thank you for your attention!

Please reach out over email if you have any questions.