

Demographic Fairness: Balance

Inspiration

Disparate Impact [[FFMSV 15](#)]

- *Griggs vs Duke Power Co.*: used non-racial features (notably, employee testing) as a proxy for race in order to discriminate against black employees in promotion
 - Duke Power Co. lost in the Supreme court, and this was deemed unlawful
 - This ruling and general philosophy helped promote affirmative action and anti-discrimination laws
 - Why is this bad:

It had a “disproportionate and adverse impact on certain individuals.”

In other words, disparate impact.

- Applied to ML: Ensure that the impact of a system across protected groups is proportionate.
- Applied to Clustering:
 - The impact on a group is measured by how many individuals of that group are in each cluster.
 - Thus, we must ensure that the number of individuals from each group in each cluster is proportional to group size

Demographic Fairness - Balance

Recall: we are given points \mathcal{C} in a metric space. We pick centers $S \subseteq \mathcal{C}$ and create a map $\varphi: \mathcal{C} \rightarrow S$. We also represent the clustering as a partitioning of points S .

Assume the points in \mathcal{C} are given colors *red* or *blue*, representing protected classes.

For a cluster C :

$$balance(C) = \min\left(\frac{\#red(C)}{\#blue(C)}, \frac{\#blue(C)}{\#red(C)}\right)$$

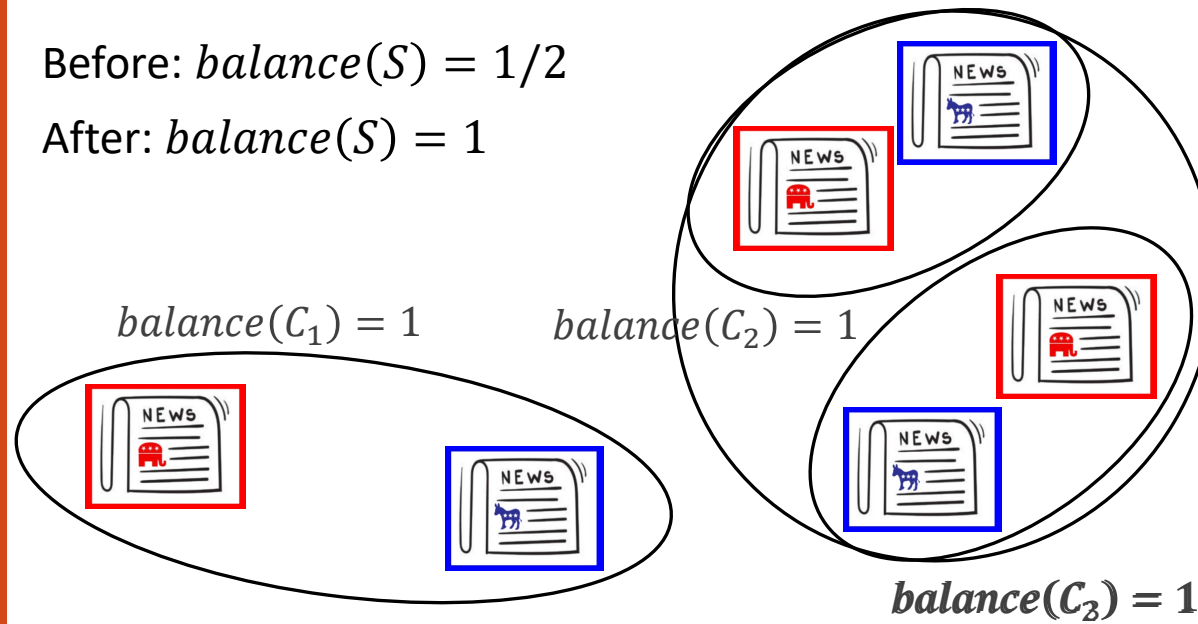
For a clustering S :

$$balance(S) = \min_{C \in S} balance(C)$$

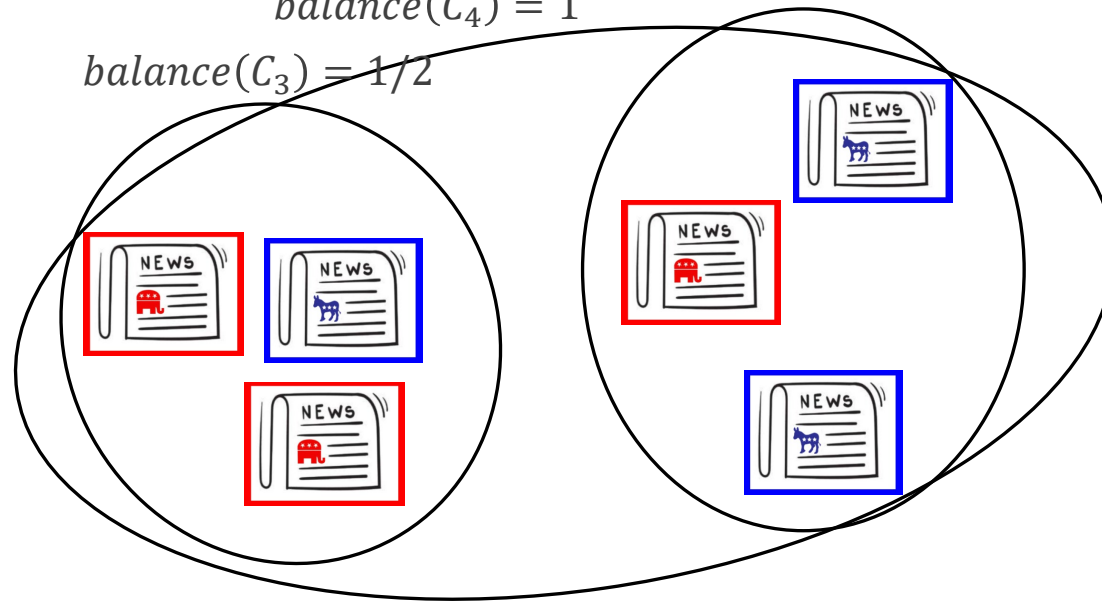
We want balance to be high (close to 1).

Before: $balance(S) = 1/2$

After: $balance(S) = 1$



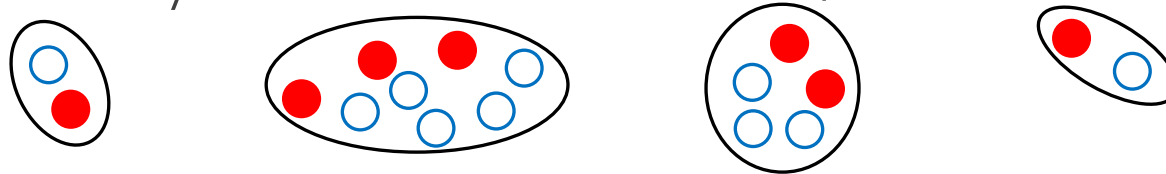
$balance(C_4) = 1$ $balance(C_3) = 1/2$ $balance(C_4) = 1/2$



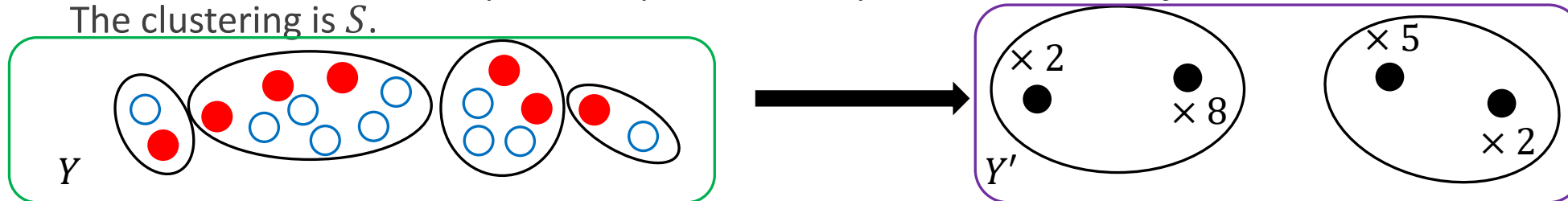
Results for Balance [CKLV 18]

Lemma: Let $\text{balance}(\mathcal{C}) \geq b/r$ for minimum integral b and r . Then we can find a clustering S with balance at least b/r and minimum cluster size $b + r$.

- 7 red, 10 blue
- $\text{balance}(\mathcal{C}) \geq 3/5$



Method: Using the previous lemma, create a *fairlet decomposition* Y , which is a fair clustering with small (but possibly too many) clusters. Run a vanilla clustering algorithm on the fairlets centers as points duplicated to equal the size of the *fairlet*, call this set Y' . The clustering is S .



Structural result: for k -median and k -center, let the objective value be ψ :

$$\psi(\mathcal{C}, S) = \psi(\mathcal{C}, Y) + \psi(Y', S)$$

Further Results for Balance [[CKLV 18](#)]

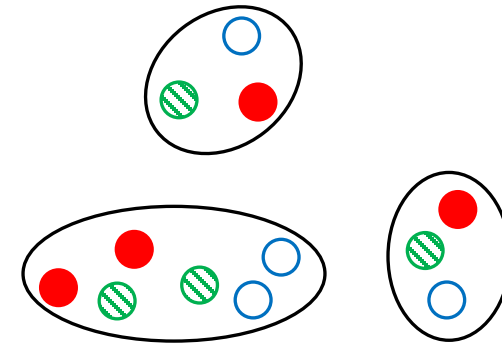
Balanced problem solved	Balance achieved	Approximation factor	Subroutine used	Subroutine approximation
1-center	1	3	1-center	2
k -center	$1/t'$	4	k -center	2
k -median	1	$2 + \sqrt{3} + \epsilon$	k -median	$1 + \sqrt{3} + \epsilon$
k -median	$1/t'$	$t' + 1 + \sqrt{3} + \epsilon$	k -median	$1 + \sqrt{3} + \epsilon$

Hardness: it is NP-hard to optimally find a $1/t'$ -balanced k -median clustering.

Fairness and Privacy [RS 18]

General fairness results

- Finds a 12-approximate fairlet decomposition on *any* number of colors
- Implies:
 - 14-approximation for k -center
 - 15-approximation for k -supplier



Fair and private clustering

- Privacy: lower bounds on the size of clusters
- Results:
 - 40-approximate private and fair k -center
 - 41-approximate private and fair k -supplier

Strongly private clustering

- Strong privacy: lower bounds on number of points of a color per cluster
- Results
 - 4-approximate strongly private k -center
 - 5-approximate strongly private k -supplier

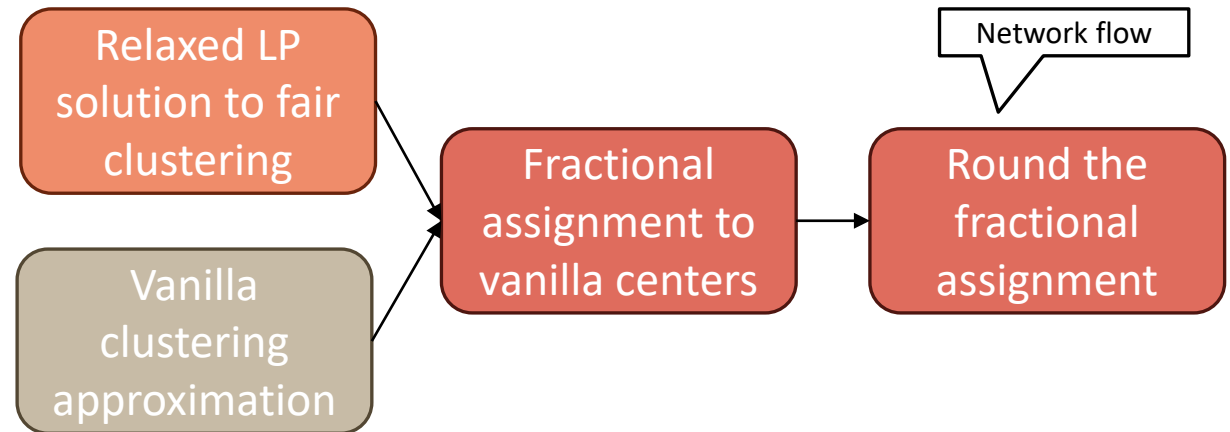
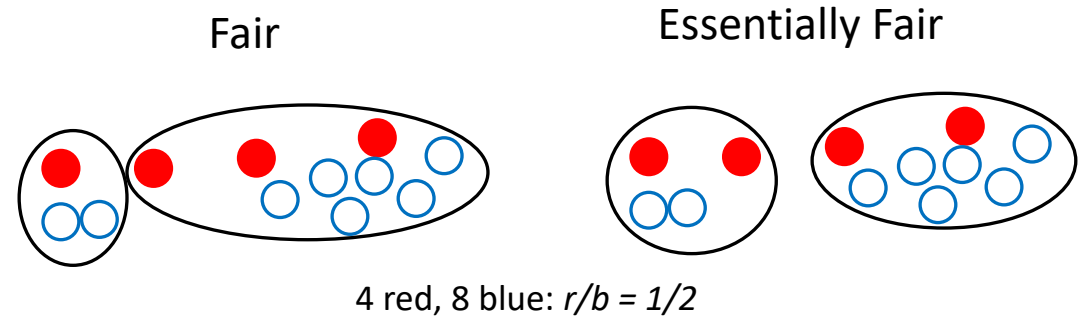
Fairness and Essential Fairness [[BGKKRSS 19](#)]

General fairness results

- 5-approximate fair k -center
- 7-approximate fair k -supplier

Essentially fair results

- Clusterings with only *additive* fairness violations:
 - E.g., you can have one extra red point in a cluster
- Results:
 - 3-approximate essentially fair k -center
 - 5-approximate essentially fair k -supplier
 - 3.488-approximate essentially fair facility location
 - 4.675-approximate essentially fair k -median
 - 62.856-approximate essentially fair k -means



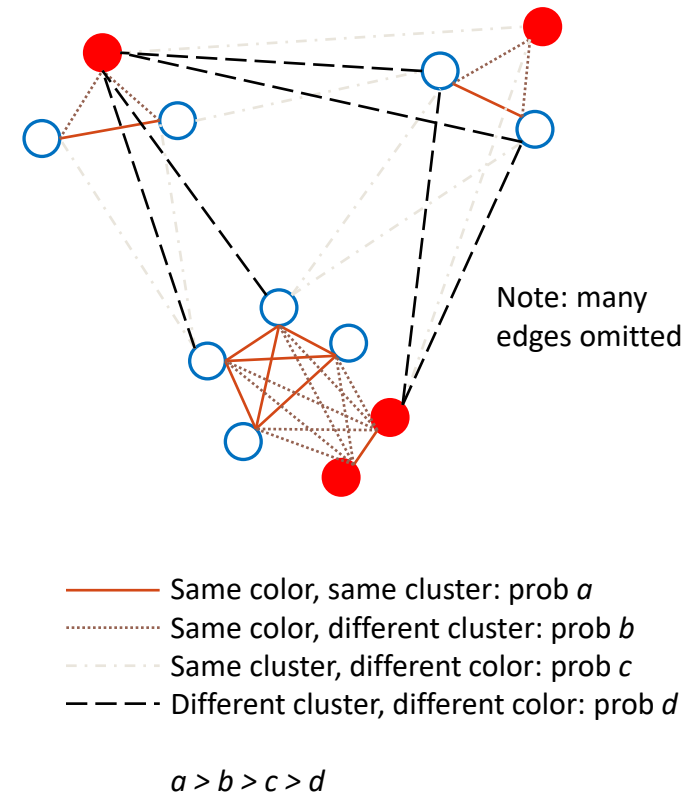
Fair Spectral Clustering [[KSAM 19](#)]

Definition – Stochastic Block Model

- There is a fair ground truth clustering
- Generate edges of weight +1 according to color and ground truth cluster

Fair spectral clustering

- Spectral clustering: create a clustering that minimizes the value of RatioCut
 - $RatioCut(S) = \sum_{C \in S} \frac{\sum_{e \in C \times V \setminus C} w(e)}{|C|}$, e.g., the sum of the ratios of the weights exiting a cluster to the size of the cluster
- Proposes a new spectral clustering algorithm:
 - Bounds the error relative to the ground truth clustering
 - Uses $O(n^3)$ time, $O(n^2)$ space



Summary - Balance

First introduced as a concept in 2018 [\[CKLV 18\]](#)

- They also developed the “fairlet decomposition” technique and came up with initial results

Many of the best approximations are from [\[BGKKRSS 19\]](#), who studied many clustering problems.

Variants explored:

- With privacy [\[RS 18\]](#)
- Spectral Clustering [\[KSAM 19\]](#)
- Essential fairness [\[BGKKRSS 19\]](#)

Demographic Fairness: Bounded Representation

Demographic Fairness — Bounded Representation [BGKKRSS 19]

Instead of requiring perfect balance, we are only constrained by given bounds. There are multiple versions studied:

α -bounded for constant α :

- Every color must represent *at most* an α fraction of any cluster

α, β -bounded, for vectors α, β :

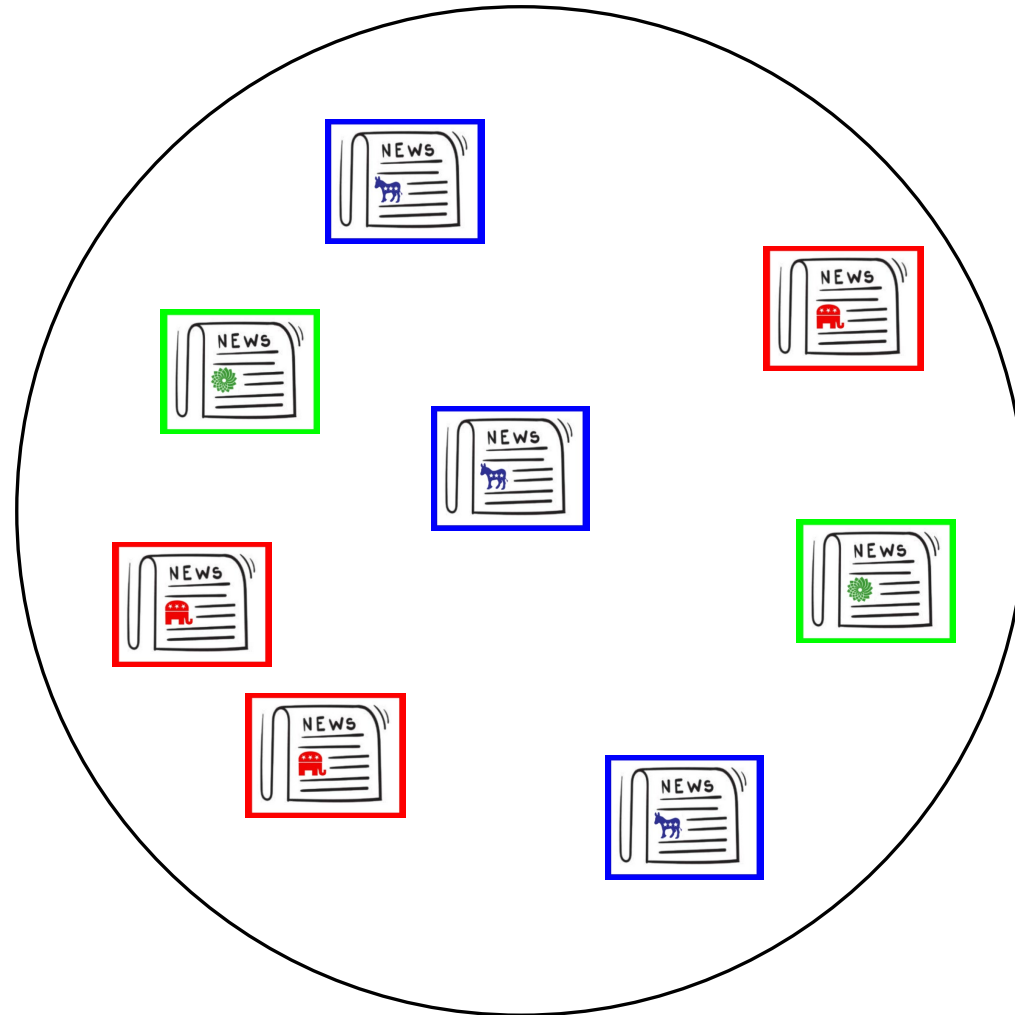
- For any color $i \in \{1, \dots, c\}$, color i must represent *at most* an α_i fraction of any cluster and *at least* a β_i fraction of any cluster

A clustering is fair if every cluster satisfies the bounded representation constraint.

Fair up to $\alpha=3/8$

Fair down to $\alpha_{\text{red}}=3/8, \alpha_{\text{blue}}=3/8, \alpha_{\text{green}}=1/4$

Fair up to $\beta_{\text{red}}=3/8, \beta_{\text{blue}}=3/8, \beta_{\text{green}}=1/4$



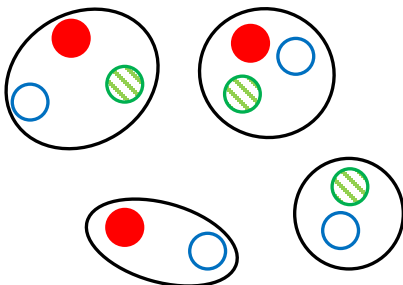
Mitigating Over-Representation [[AEKM 19](#)]

Fairness constraint: general lower bound α

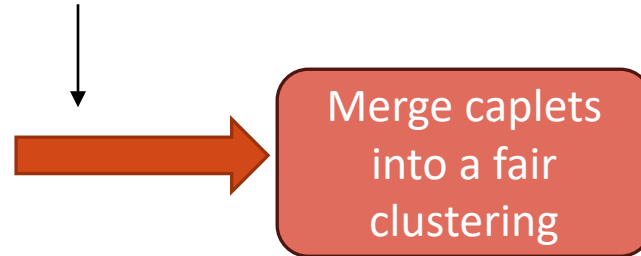
Important technique: create a threshold graph on bichromatic edges of weight $\leq \tau$

Results

- General α : 3-approximation with violation 2
- $\alpha=1/t$: 3-approximation with violation 1
- $\alpha=1/2$: 12-approximation with *no* violation



Caplet decomp. on threshold graph (exists)



Solve an LP relaxation on weights $\leq 3\text{OPT}$

Filter facilities: only select distant ones

Round to an integral solution

Here we lose some approximation factor and fairness

Fair Correlation Clustering [[AEKM 20](#)]

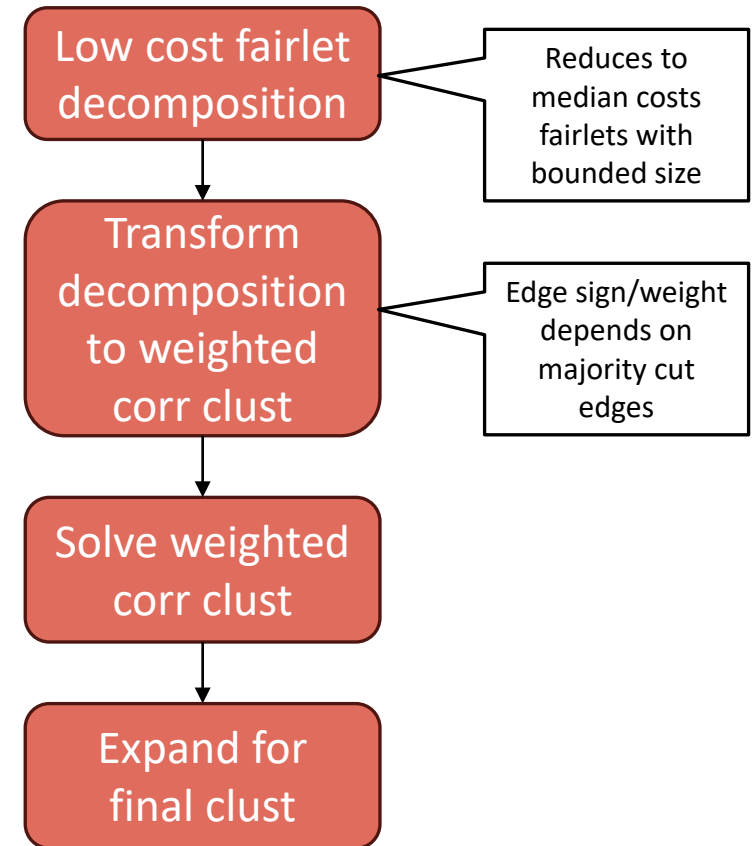
Fairness constraint: general lower bound α (also generalizes further)

Definition – correlation clustering

- Edges are all +1 or -1
- Minimize: the (weighted) sum of the +1 edges between clusters and -1 edges within clusters

Fair correlation clustering (on c colors)

- $\alpha=1/2$: 256-approximation
- $\alpha=1/c$: $(16.48c^2)$ -approximation
- $\alpha=1/t$: $O(tp)$ -approximation for p -approximate median cost fairlet decomposition



Fair Hierarchical Clustering [AEKKMMPVW 20]

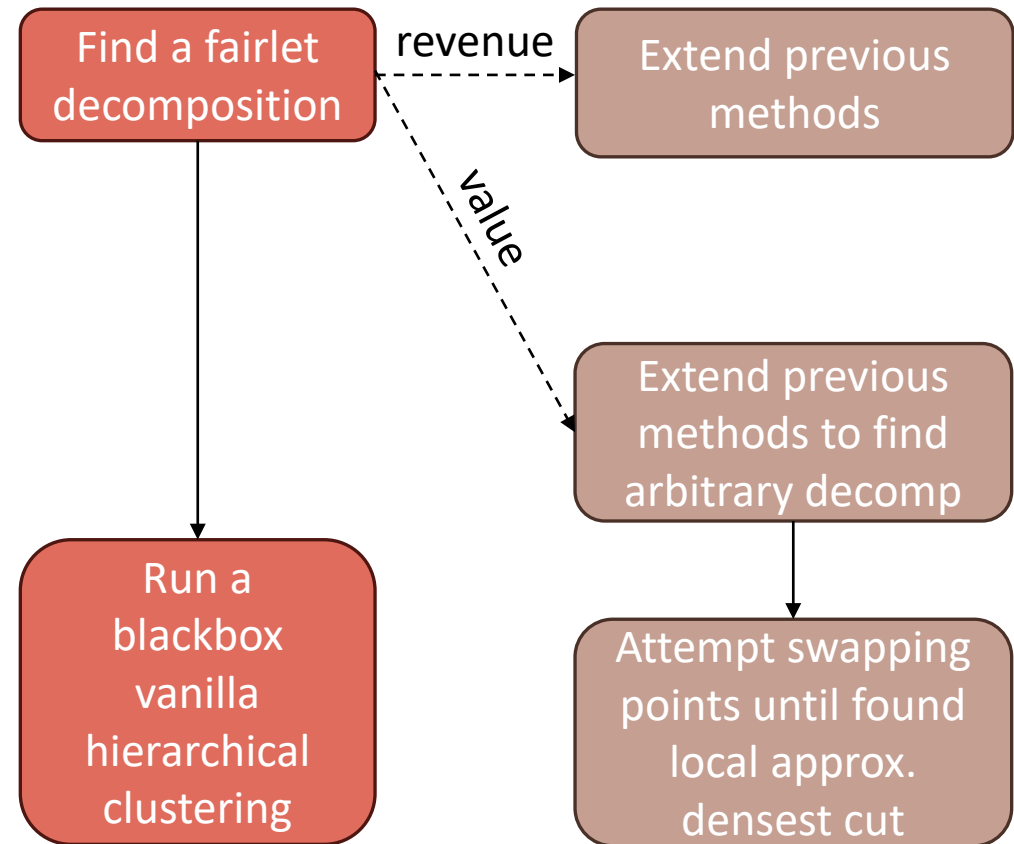
Fairness constraint: general lower bound α (also generalizes further)

Hierarchical clustering objectives

- Cost: initial objective, APX-hard
- Revenue: dual to cost, const-approximable
- Value: cost variant, const-approximable

Results

- Cost: $O(n^{5/6} \log^{5/4} n)$ -approximation (highly combinatorial methods)
- Revenue: $(1/3 - o(1))$ -approximation if $\alpha = 1/t$ or 2 colors
- Value: $(2/3 - \epsilon)(1 - o(1))$ -approximation



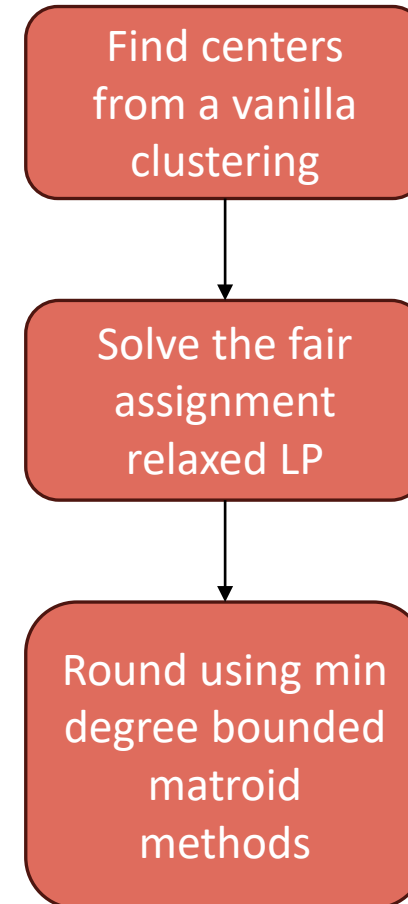
General Bounds, Overlapping Groups [[BCFN 19](#)]

Fairness constraint: specific lower and upper bound vectors α, β , also vertices can be in *multiple* groups (bounded by Δ)

Fair assignment problem: given a set of centers, what is the best way to create a fair clustering by assigning points to the centers?

Results

- Given a p -approximate vanilla k -clustering for the p -norm, gives a $p+2$ approximation with additive fairness violation $4\Delta+3$



Probabilistic Fair Clustering [[EBTD 20](#)]

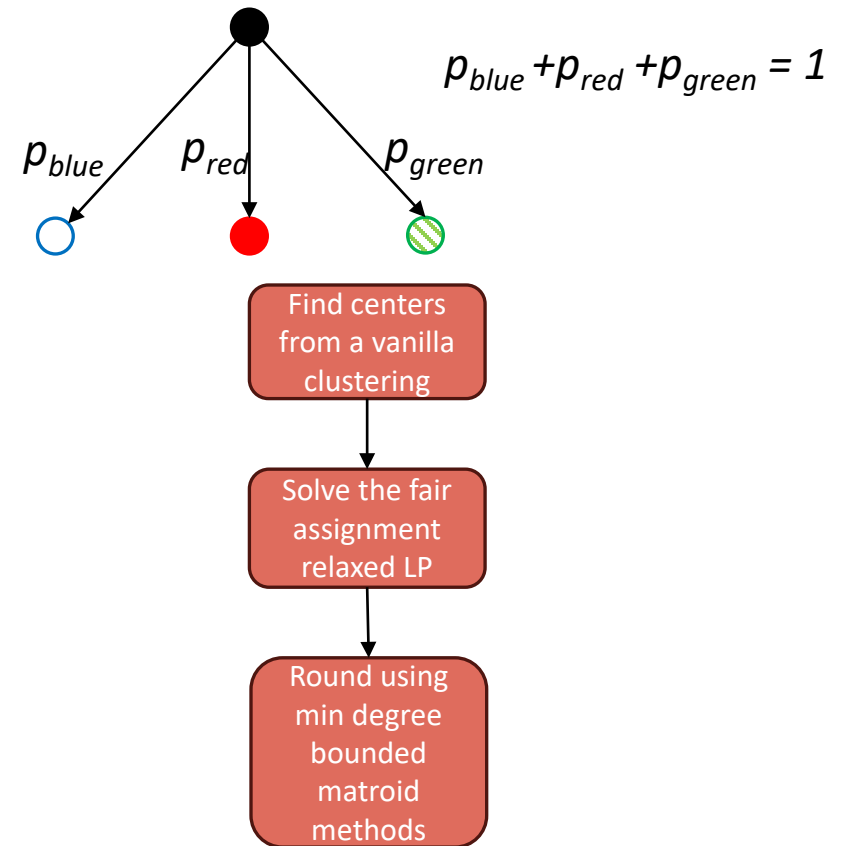
Fairness constraint: specific lower and upper bound vectors α, β , also vertices can be in *multiple* groups (bounded by Δ)

Probabilistic setting

- Colors are not given. Each point has a probability of being assigned some color
- We guarantee that the *expected number* of each color in each cluster is bounded above/below
- Addresses any p -norm

Results

- Given a vanilla p -approximation, there is a $p+2$ -approximation with +1 violation (2 colors)
- Given a vanilla p -approximation, there is a $p+2$ -approximation with +1 violation (FPT, large clusters)



Fixing a Bounded Cost [EBSD 21]

Fairness constraint: specific lower and upper bound vectors α, β , also vertices can be in *multiple* groups (bounded by Δ)

Fair clustering under bounded cost

- Fix an upper bound for the clustering cost
- Minimize the degree of unfairness for any color (i.e., the proportional violation Δ of upper and lower bounds)
 - Utilitarian: minimize the sum of Δ s
 - Egalitarian: minimize the maximum Δ
 - Leximin: minimize the maximum Δ , then second largest Δ , ...

Results

- Fair clustering (or assignment) under bounded cost is NP-hard
- Given a vanilla approximation, there are approximations for the fair bounded cost problem

Normal linear program

Minimize: $cost(S)$

Subject to: $\alpha_i |C| \leq |i(C)| \leq \beta_i |C|$

New linear program

Minimize: $\sum \Delta$ or $\max \Delta$ or $\max \max \dots$

Subject to: $cost(S) \leq upper\ bound$

and: $(\alpha_i + \Delta) |C| \leq |i(C)| \leq (\beta_i + \Delta) |C|$

Summary: Bounded Representation

Two versions:

- α -capped clustering
 - Solved with very little to no additive fairness violation [[AEKM 19](#)]
- α, β -bounded, with upper and lower bound vectors
 - Solved with additive fairness violation [[BCFN 19](#)]

Variants explored:

- α -capped correlation clustering
- α -capped hierarchical clustering
- α, β -bounded probabilistic clustering
- α, β -bounded clustering with bounded cost

However, both are stated
in terms of “union closed”
constraints

Demographic Fairness: Bounds on Chosen Centers

Demographic Fairness — Bounds on Chosen Centers

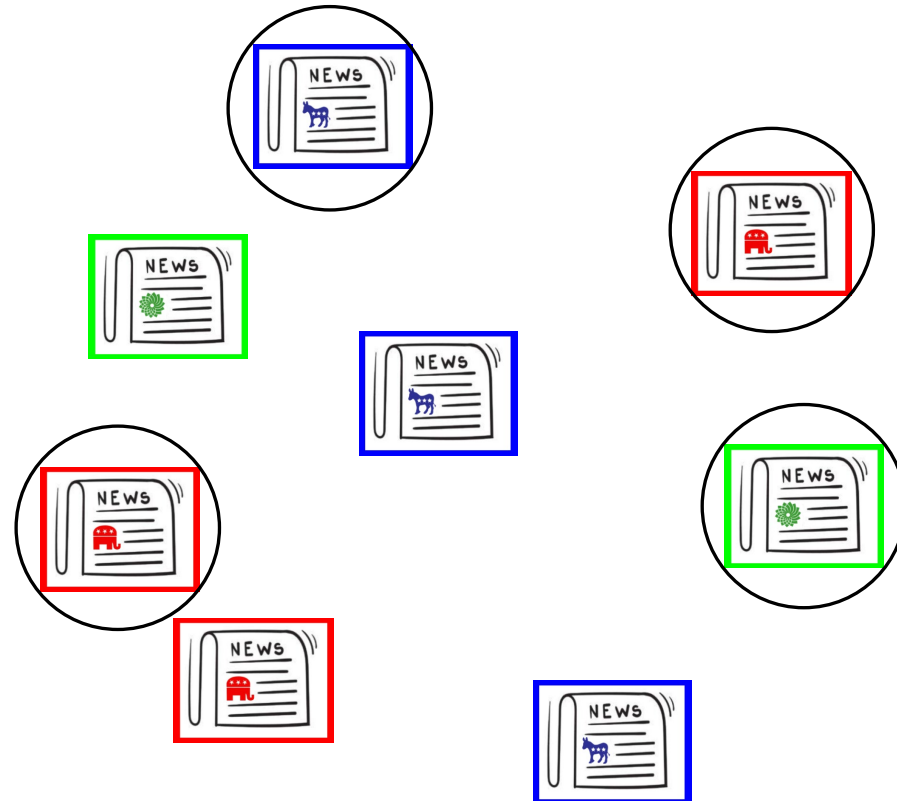
Data summarization: do a k -clustering. The resulting centers are then outputted as *representatives* of the data set.

Fairness: for every color i , there must be at least k_i centers of color i .

Let $k_{red} = k_{blue} = k_{green} = 1$

This clustering is not fair.

However this alternate clustering is fair



Data Summarization [[KAM 19](#), [JNN 20](#)]

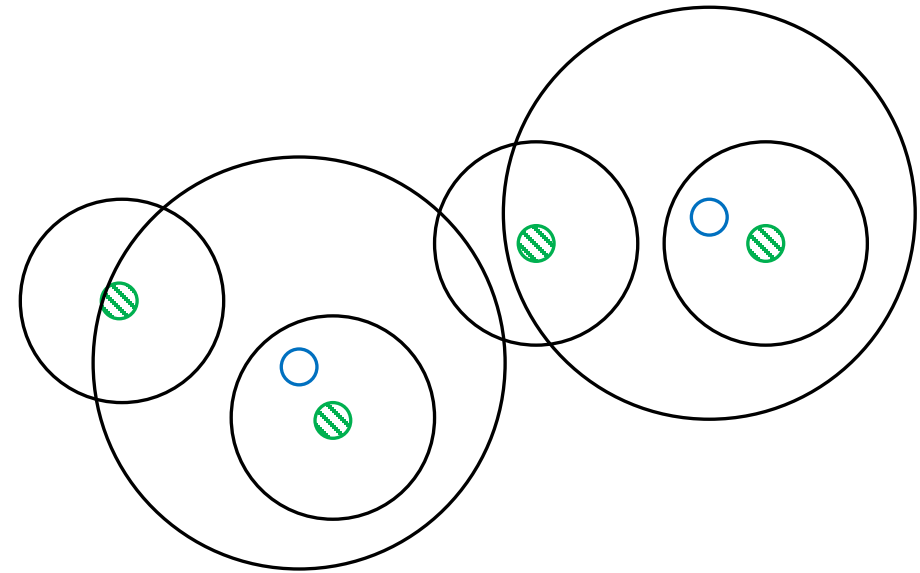
First introduced the fairness with regards to bounds on chosen centers.

Results [[KAM 19](#)]

- Fair data summarization for k -center has a 5-approximation on 2 colors (tight) that runs in time $O(kn)$
- Fair data summarization for k -center has a $(3 \times 2^{c-1} - 1)$ approximation on c colors that runs in time $O(kc^2n + kc^4)$

Results [[JNN 20](#)]

- Fair data summarization for k -center has a 3-approximation on c colors (tight) that runs in time $O(kn)$



Diversity-Aware k -Means [TOG 21]

		NP-Hard?	FPT(k)?	Approx factor	Approx method
General case		Yes	No	X	X
c colors	$\sum_{i \in [c]} k_i = k$	Yes	?	8	LP
c colors	$\sum_{i \in [c]} k_i < k$	Yes	?	8	$O(k^{c-1})$ LP calls
2 colors	$k_1 + k_2 = k$	Yes	?	$3+\epsilon$	Local search
2 colors	$k_1 + k_2 < k$	Yes	?	$3+\epsilon$	$O(k)$ local search calls

Demographic Fairness: Proportionality

Demographic Fairness — Proportionality

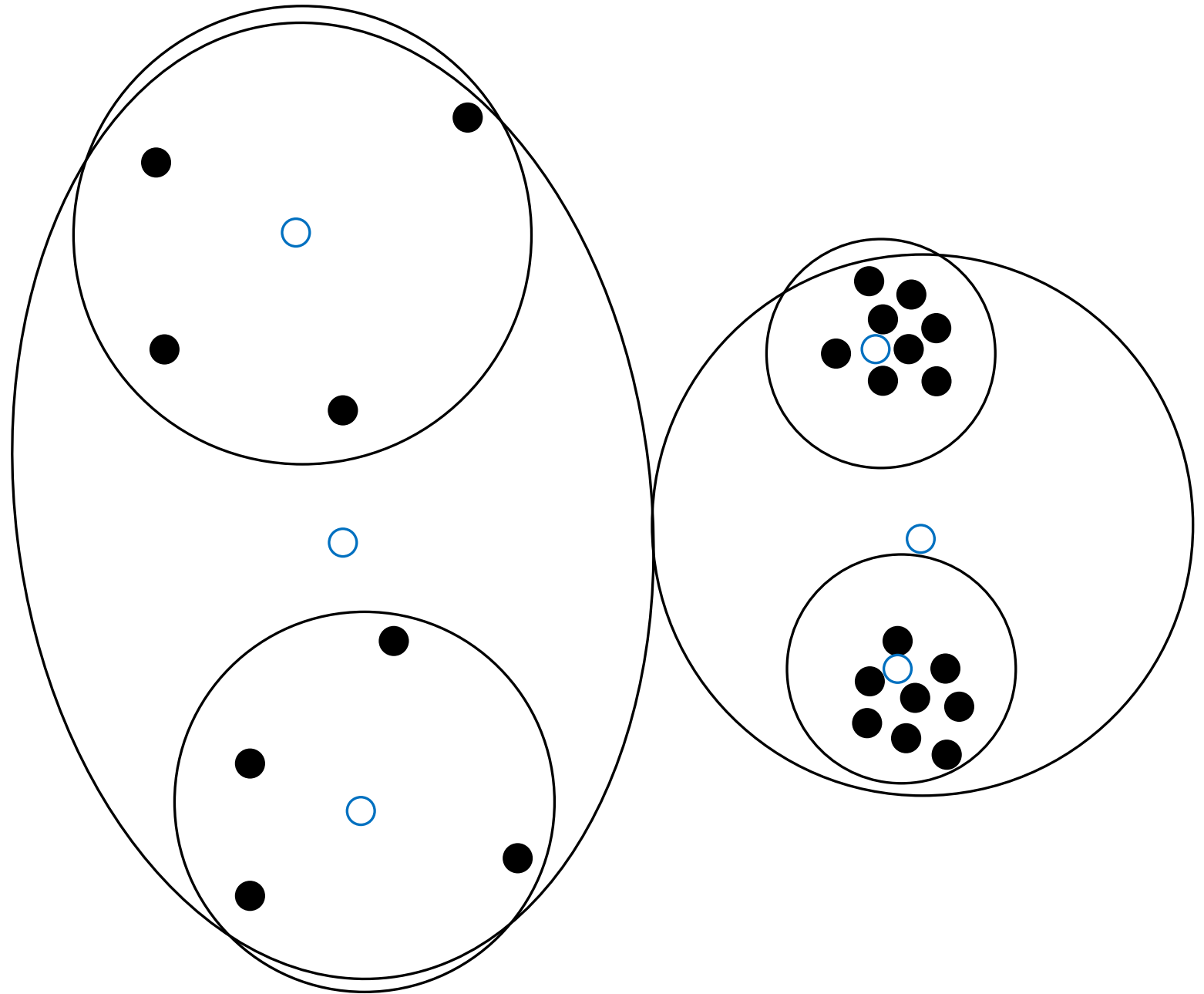
Idea: Every set of at least n/k points is entitled to its own cluster.

Blocking coalition: a set of pn/k points such that we can add a center that is closer to all points in the set than their assigned center.

- “ p -proportional”

Benefits

- Pareto optimality: Let X and X' be two proportional solutions. Then there is some point that X “treats” at least as well as X' .
- Oblivious: Independent of sensitive attributes.
- Robust: Outliers cannot form coalitions.
- Scale invariant



Proportionally Fair Clustering [[CFLM 19](#)]

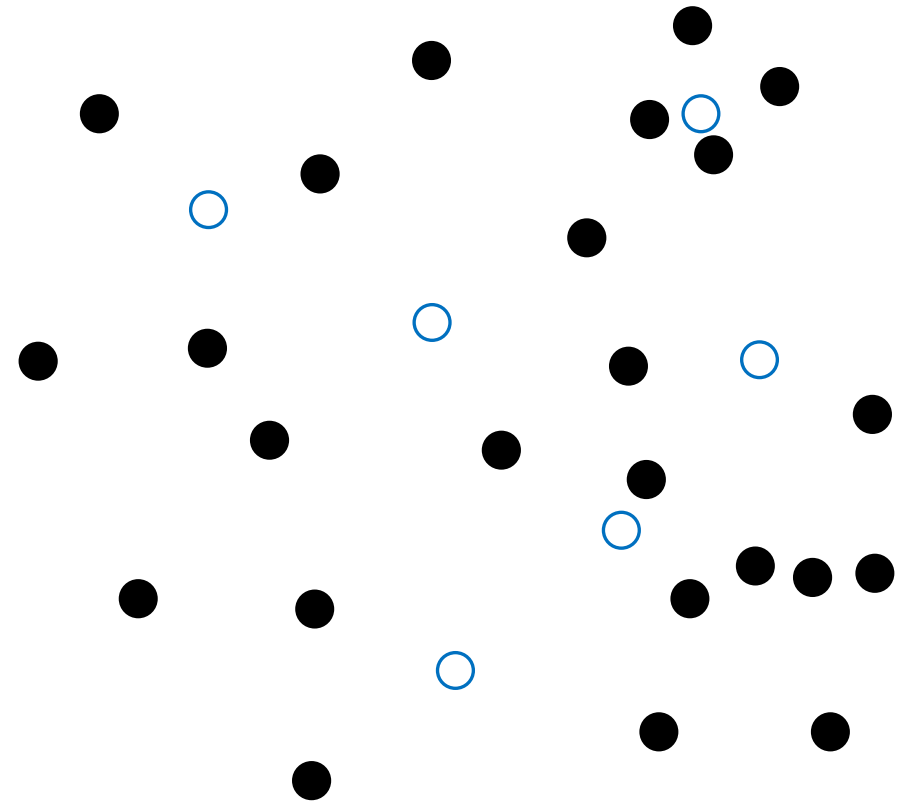
Introduced the problem of proportionally fair clusterings.

Hardness

- There may be no 2-proportional solution

Results

- $(1 + \sqrt{2})$ -proportional solution
- $O(1)$ -proportional solution 8-approximates k -median
- Uniform random sampling approximately preserves the proportionality of any set of centers w.h.p.
- Good heuristic local search algorithm that finds nearly proportional solutions



More Proportionally Fair Clustering [[MS 20](#)]

Considers [[CFLM 19](#)] in metric spaces defined by different norms.

Results

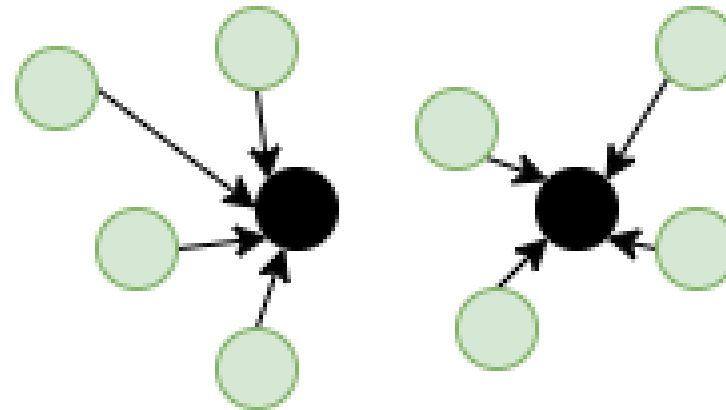
- $(1 + \sqrt{2})$ general approximation is really a 2-approximation for L^2
- Shows tightness of $(1 + \sqrt{2})$ approximation for L^1 and L^∞
- In L^2 , we cannot do better than $2/\sqrt{3}$
- In L^1 and L^∞ , we cannot do better than 1.4
- Using tree distance or graph distance when $k \geq n/2$, exact proportionality exists
- In L^2 and many dimensions, checking existence is NP-hard, and the original algorithm is only in NP (it is a PTAS in the dimensionality)
- When there are infinitely many centers, proportionality preservation under random sampling holds even in L^2 and many dimensions

(End slide)

Demographically Fair Clustering with Outliers

k-Clustering with Outliers

- Points \mathcal{C} in metric space with distance $d: \mathcal{C}^2 \rightarrow \mathbb{R}_{\geq 0}$
- Pick $S \subseteq \mathcal{C}$ with $|S| \leq k$
- Pick $\mathcal{A} \subseteq \mathcal{C}$ with $|\mathcal{A}| \geq m$
- Construct $\varphi: \mathcal{A} \rightarrow S$ such that some objective is minimized
- Allowed to exclude a certain number of points from the optimization objective:
 - ❖ Robustness: Avoid noise in the data
 - ❖ Scarce resources: Servicing only a certain fraction of the population is acceptable



Motivational Examples

➤ **Clustering Setting:**

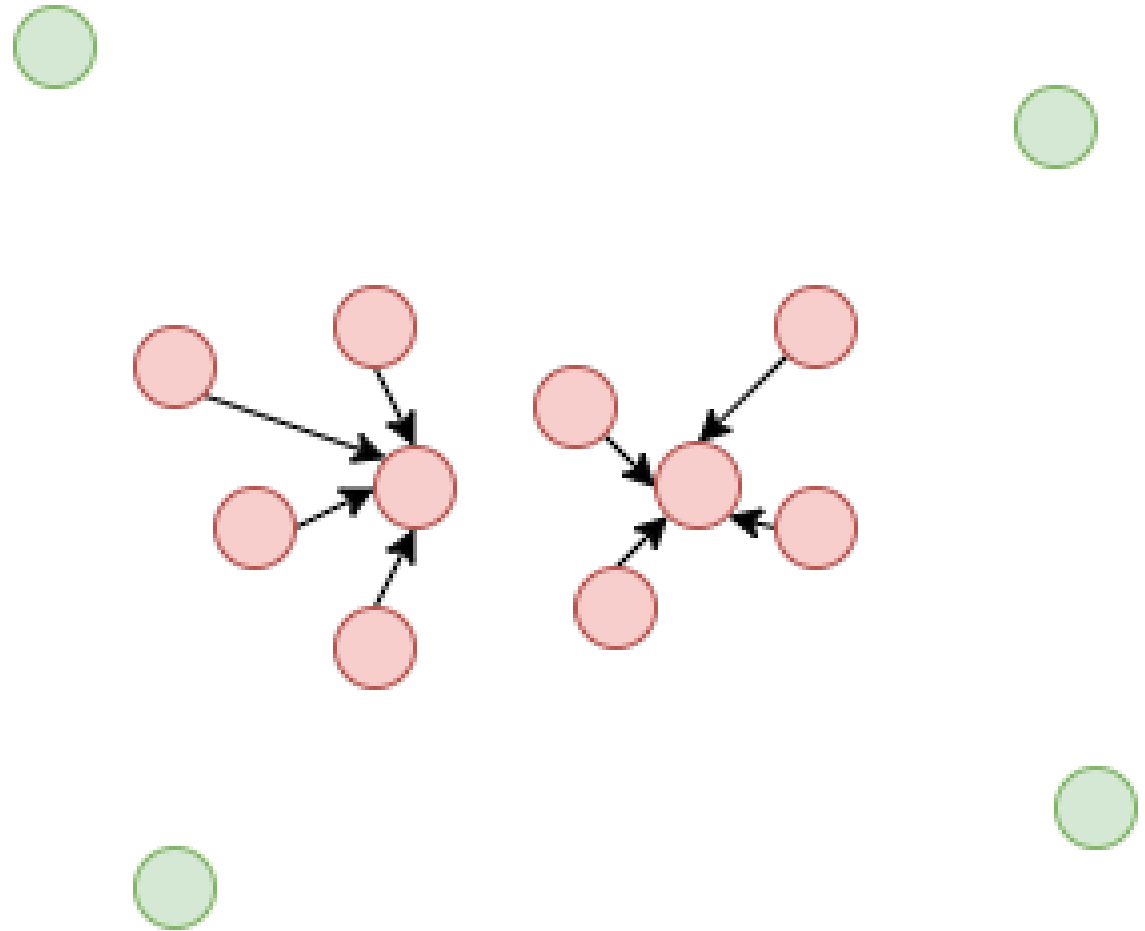
- ❖ The points are users of a website.
- ❖ The website wants to cluster its users in groups of high similarity, so that it offers relevant recommendations.
- ❖ The cluster center is thought of as the most representative point of the cluster.
- ❖ Points with unique profiles might be excluded.

➤ **Facility Location Setting:**

- ❖ Points correspond to cities/towns/counties.
- ❖ A state wants to place vaccination sites in a metric space.
- ❖ Each point should have a vaccination center in close vicinity.
- ❖ Due to scarce resources, it may be acceptable to provide a good covering guarantee to only a fraction of the population.

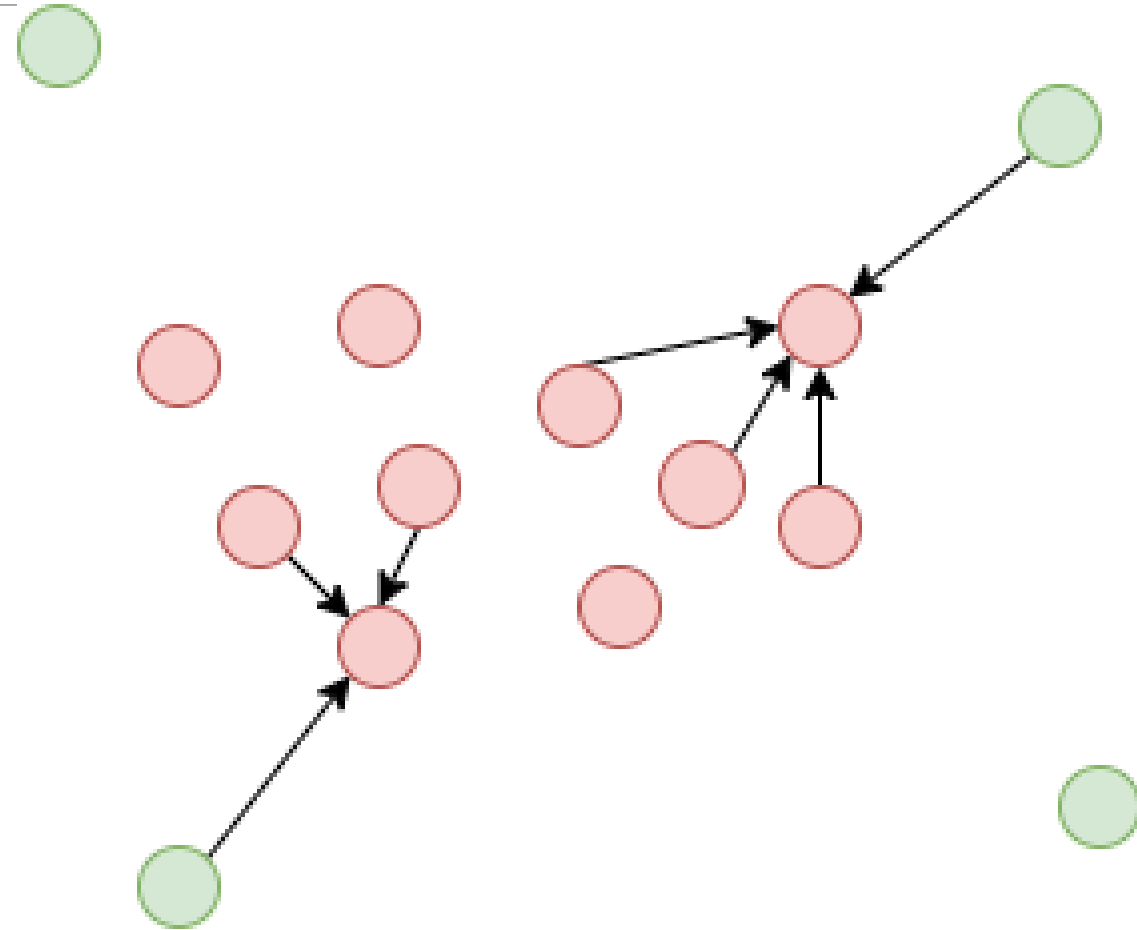
Bias in Clustering with Outliers

- **Being an outlier is disadvantageous!!!**
 - Example 1: Outliers will not receive any recommendations
 - Example 2: Outliers will not enjoy close access to a vaccination center
- Suppose the points of \mathcal{C} come from γ demographic groups $\mathcal{C}_1, \dots, \mathcal{C}_\gamma$.
- A solution can be biased if it disproportionately views points from certain groups as outliers.



Fair Clustering with Outliers

- Proposed fix:
 - ❖ For each group \mathcal{C}_l we are given a value $m_l \geq 0$
 - ❖ Instead of $|\mathcal{A}| \geq m$, we now require $|\mathcal{A} \cap \mathcal{C}_l| \geq m_l$ for every $l \in [\gamma]$
- Example with $m_{red} = 5$ and $m_{green} = 2$
- Arbitrary m_l values can capture a plethora of fairness scenarios
 - ❖ Equitable treatment, e.g., $m_l \geq |\mathcal{C}_l|/2$ for all $l \in [\gamma]$
 - ❖ Preferential treatment, e.g., give a higher coverage guarantee to demographics that really need it



Results

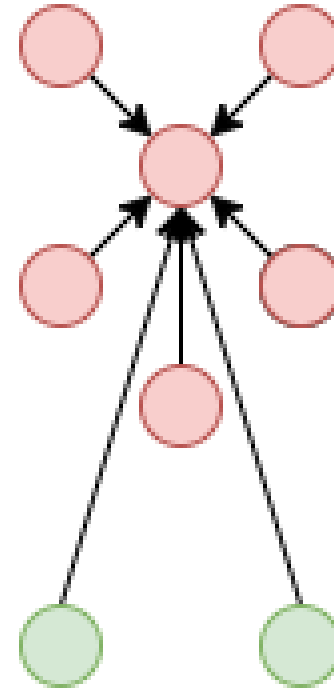
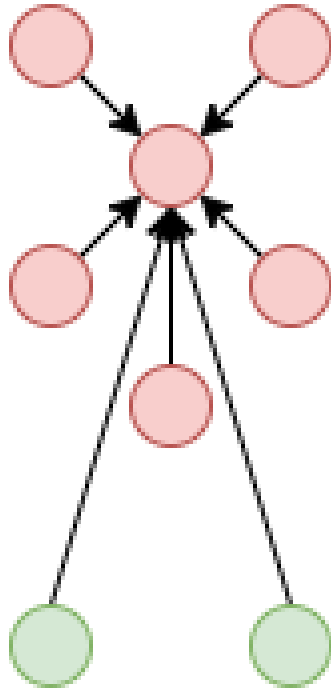
- The problem has only been studied for the k-center objective, i.e., minimize $\max_{j \in \mathcal{C}} d(j, \varphi(j))$, under the name Fair Colorful k-Center
- It was introduced by Bandyapadhyay et al. (“A Constant Approximation for Colorful k-Center”- ESA 2019), who gave a 17-approximation algorithm for it in the Euclidean plane, when $\gamma = O(1)$.
- Anegg et al. (“A Technique for Obtaining True Approximations for k-Center with Covering Constraints”) and Jia et al. (“Fair Colorful k-Center Clustering”) independently gave a 4-approximation and a 3-approximation respectively, both appearing in IPCO 2020.
- Both of the above algorithms work for general metrics, when $\gamma = O(1)$.
- Anegg et al. also showed that when γ is not a constant, there cannot exist any non-trivial approximation for the problem, unless P=NP.

Socially Fair k-Clustering

Motivation

- In many clustering or facility location applications the quantity is $d(j, \varphi(j))$ (referred to as “assignment distance”) is what really matters.
 - ❖ Clustering: It measures how representative $\varphi(j)$ is for j .
 - ❖ Facility Location: It represents the distance j needs to travel in order to reach its service provider $\varphi(j)$.
- The smaller $d(j, \varphi(j))$ is the more satisfied the point j .
 - 1) Recall the previously mentioned recommendation system application.
 - 2) Recall the previously mentioned vaccination sites allocation application.
- Conclusion: If \mathcal{C} consists of γ demographic groups $\mathcal{C}_1, \dots, \mathcal{C}_\gamma$, then we should be fair in terms of the assignment distances provided to the points of different groups.

Example of a Biased Solution



Results

- The problem was introduced independently by Ghadiri et al (“Socially Fair k -Means Clustering”) and Abbasi et al. (“Fair Clustering via Equitable Group Representations”) at FAccT 2021.
- Both papers demonstrated an $O(\gamma)$ -approximation algorithm.
- Makarychev and Vakilian (“Approximation Algorithms for Socially Fair Clustering” – COLT 2021) gave an $O(\frac{\log \gamma}{\log \log \gamma})$ -approximation algorithm. They also showed that this is the best possible approximation ratio for the problem.
- Goyal and Jaiswal (“Tight FPT Approximation for Socially Fair Clustering” – Arxiv 2021) give a tight $(3+\epsilon)$ -approximation algorithm for the problem, that runs in FPT time of $\left(\frac{k}{\epsilon}\right)^k$.