Fair Representation Clustering

Instructor: Yury Makarychev, TTIC

Max Planck ADFOCS 2023

k-means and k-medians

Given a set of points X in a metric space

Partition X into k clusters $C_1, \ldots, C_k$ and find a center c_i for each C_i so as to minimize

$$\sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i) \qquad (k\text{-medians})$$

$$\sum_{i=1}^{k} \sum_{u \in C_i} d(u, c_i)^2 \qquad (k\text{-means})$$

Fair representation clustering

Given:

- a set of points X and a distance function d on X.
- each point belongs to one of the groups $G_1, \ldots, G_\ell \subset X$
- parameters α_i and β_i for each group i

Fair representation:

A clustering $C_1, \ldots, C_k$ has fair representation if

$$\alpha_j |C_i| \le |C_i \cap G_j| \le \beta_j |C_i|$$

Goal: Find a fair representation clustering that minimizes the k-median or k-means objective.

Plan for Today

Equal representation

[Chierichetti, Kumar, Lattanzi, Vassilvitskii ‘17] $\ell = 2$

[Böhm, Fazzone, Leonardi, Schwiegelshohn ‘21] $\ell > 2$

General representation requirements (pseudoapproximation with a constant additive violation)

[Bera, Chakrabarty, Flores, and Negahbani ‘19]
[Bercea, Groß, Khuller, Kumar, Rösner, Schmidt, and Schmidt ‘18

True approximation for constant ℓ

[Dai, M, Vakilian ‘22]

Equal Representation

Consider an important special case when

$$\alpha_j = \beta_j = 1/\ell$$

We want to ensure that each cluster has the same number of points from each group.

Important: $C_1, \dots, C_k$ don't have to be Voronoi clusters

Based on

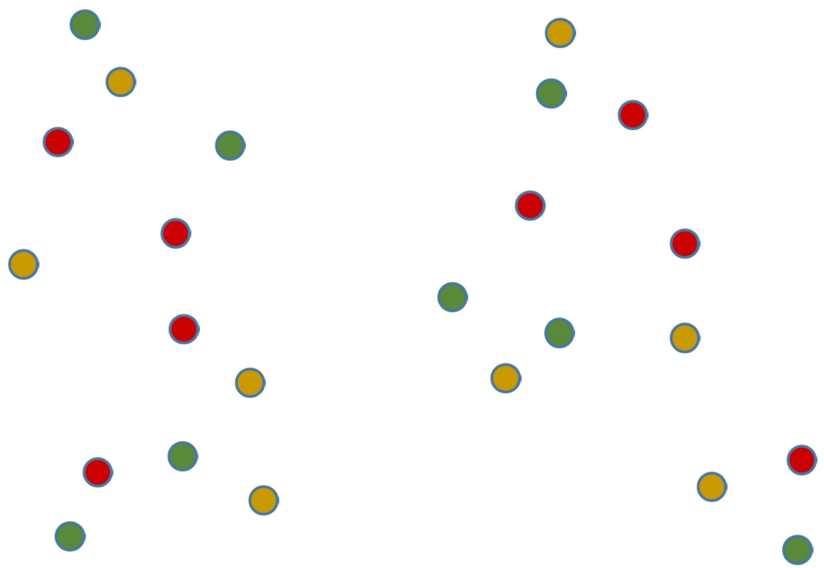
[Chierichetti, Kumar, Lattanzi, Vassilvitskii '17] $\ell = 2$

[Böhm, Fazzone, Leonardi, Schwiegelshohn '21] $\ell > 2$

Idea [Chierichetti et al]

Partition the dataset into $t = n/\ell$ sets $F_1, \ldots, F_t$ called fairlets such that each F_i has fair representation:

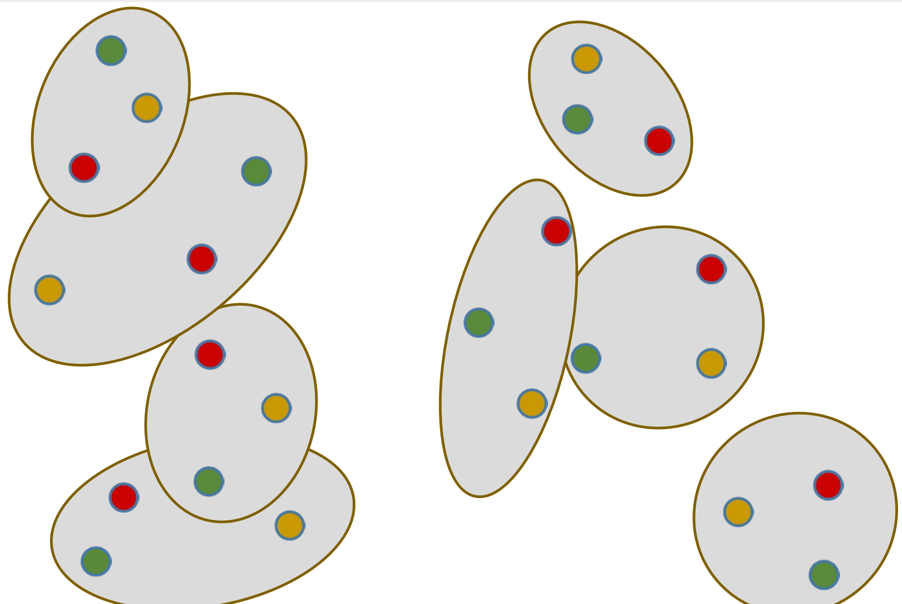
F_i contains exactly one point from each G_i



Idea [Chierichetti et al]

Partition the dataset into $t = n/\ell$ sets $F_1, \dots, F_t$ called fairlets such that each F_i has fair representation:

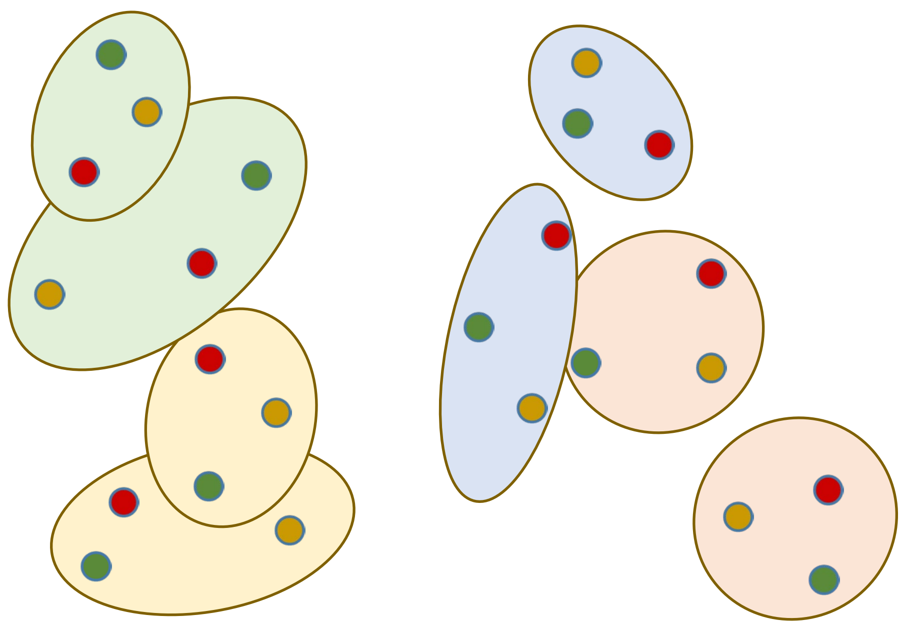
F_i contains exactly one point from each G_i



Idea [Chierichetti et al]

F_i contains exactly one point from each G_i

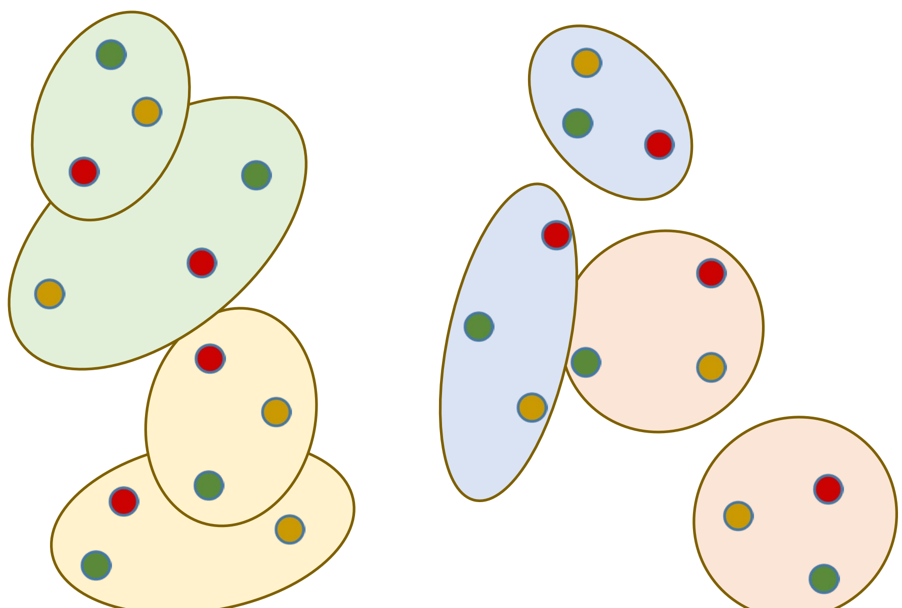
Assign fairlets to clusters $C_1, \ldots, C_k$.



Why?

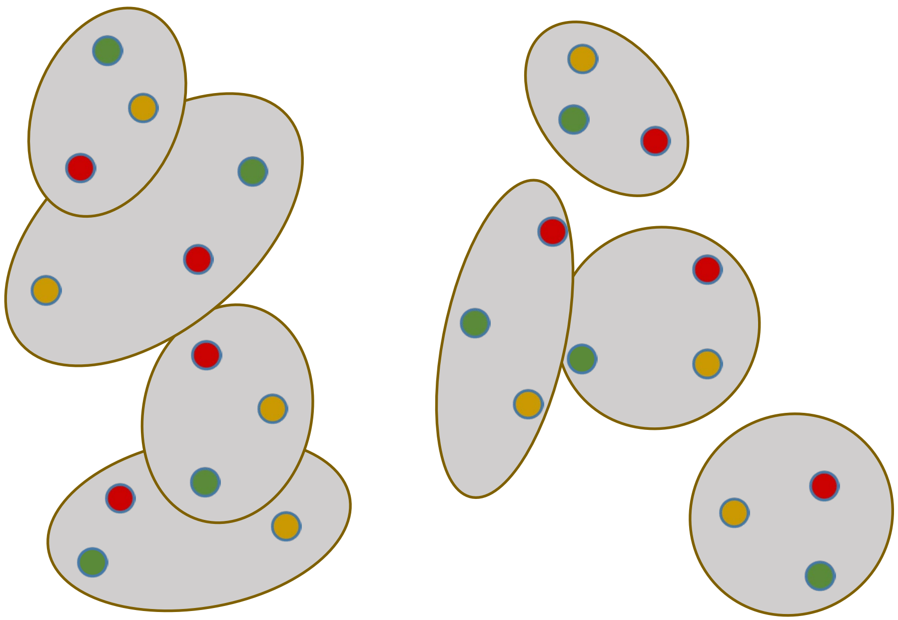
! The obtained clustering has fair representation

It’s possible to obtain the optimal solution in this way.



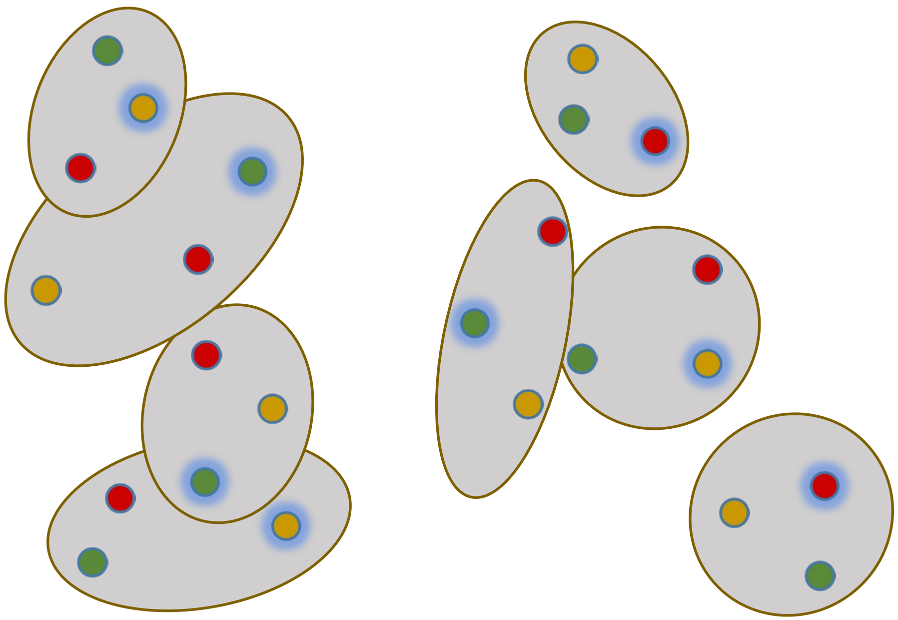
Meta-algorithm

- Find a fairlet decomposition
- Choose a representative y_i in each fairlet F_i
- Run a standard algorithm to cluster $Y = \{y_1, \ldots, y_t\}$



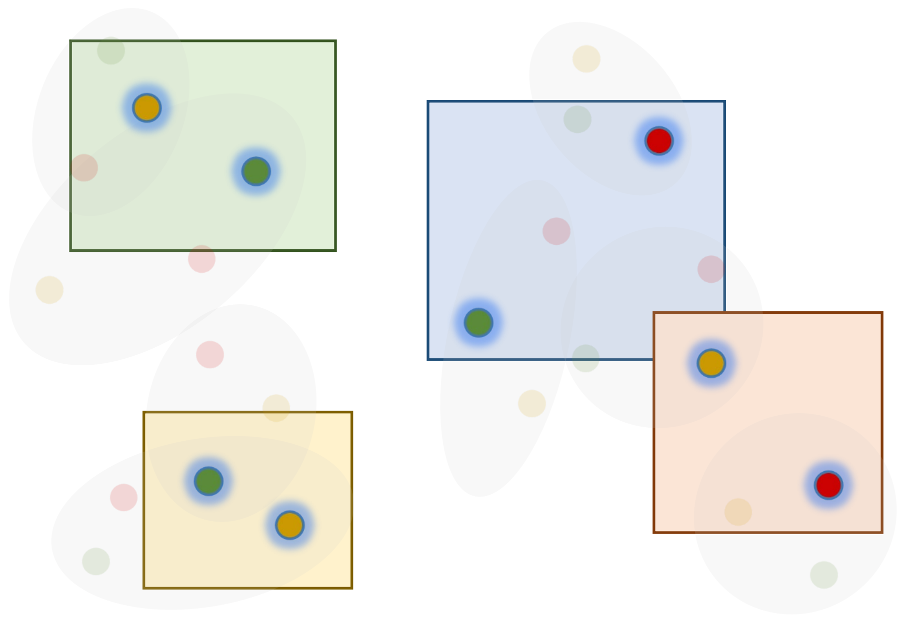
Meta-algorithm

- Find a fairlet decomposition
- Choose a representative y_i in each fairlet F_i
- Run a standard algorithm to cluster $Y = \{y_1, \dots, y_t\}$



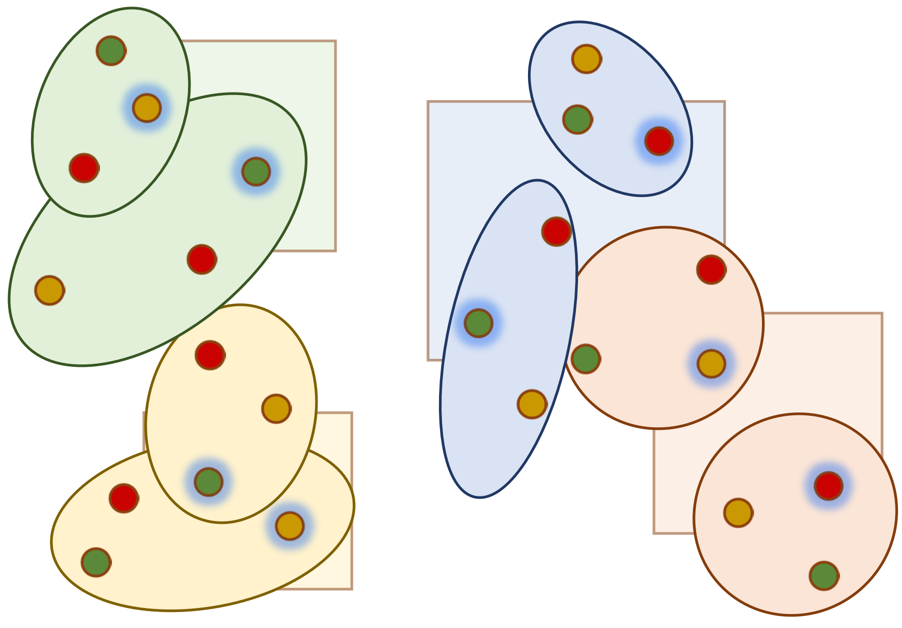
Meta-algorithm

- Find a fairlet decomposition
- Choose a representative y_i in each fairlet F_i
- Run a standard algorithm to cluster $Y = \{y_1, \dots, y_t\}$



Meta-algorithm

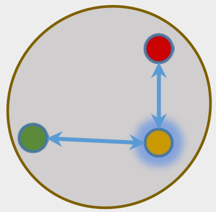
- Find a fairlet decomposition
- Choose a representative y_i in each fairlet F_i
- Run a standard algorithm to cluster $Y = \{y_1, \ldots, y_t\}$



Core Lemmas and Definitions

Define the cost of the fairlet decomposition F_i with representatives $Y = \{y_i\}$ as:

$$\mathrm{fairlet}(F, Y) = \sum_{i=1}^{t} \sum_{u \in F_i} d(u, y_i)$$



The cost of clustering $C_1, \dots, C_k$ with centers $c_1, \dots, c_k$ of a set S is

$$\mathrm{cost}(S, C_i, c_i) \equiv \mathrm{cost}(S) = \sum_{i} \sum_{u \in C_i \cap Y} d(u, c_i)$$

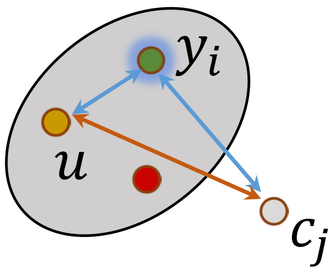
Core Lemmas and Definitions

Lemma 1: Consider a fairlet decomposition F_i with representatives Y. Let $C_1, \ldots, C_k$ be a clustering of Y with centers c_i. Extend it to X. Then

$$\mathrm{cost}(X) \leq \ell \cdot \mathrm{cost}(Y) + \mathrm{fairlet}(F, Y)$$

Proof: for $u \in F_i \subseteq C_j$

$$d(u, c_j) \leq d(u, y_i) + d(y_i, c_j)$$



C_i^* is the optimal fair clustering

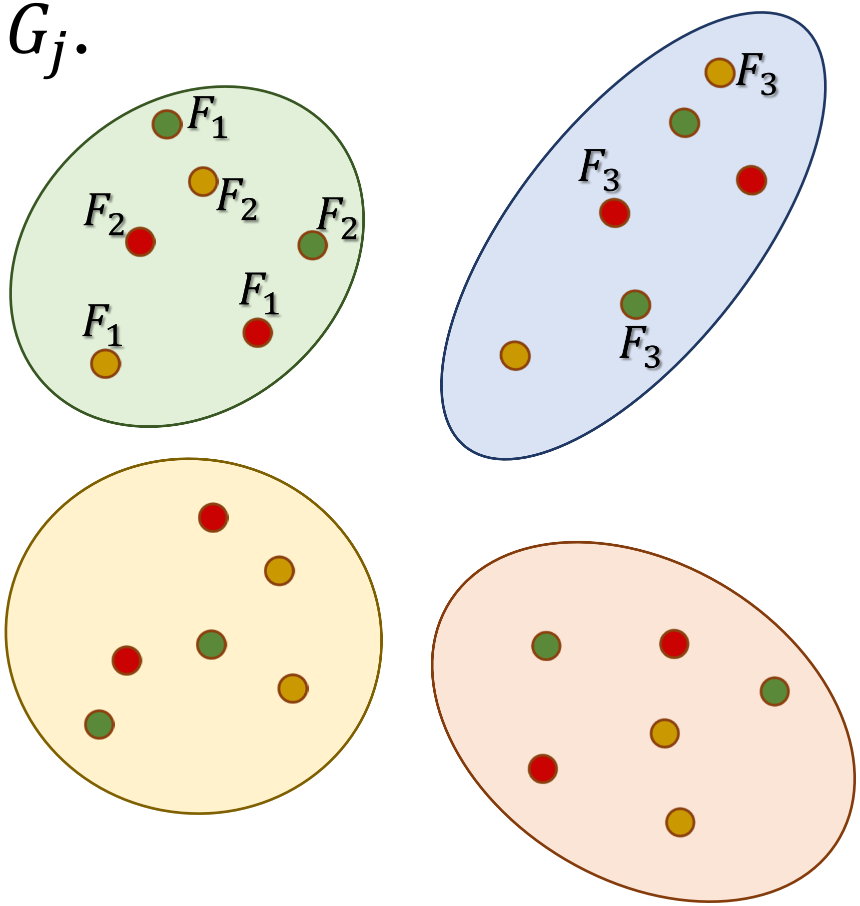
Core Lemmas and Definitions

Lemma 2: There exists a fairlet decomposition $F_1, \dots, F_t$ with representatives $Y = G_{j^*}$ for some j^* s.t.

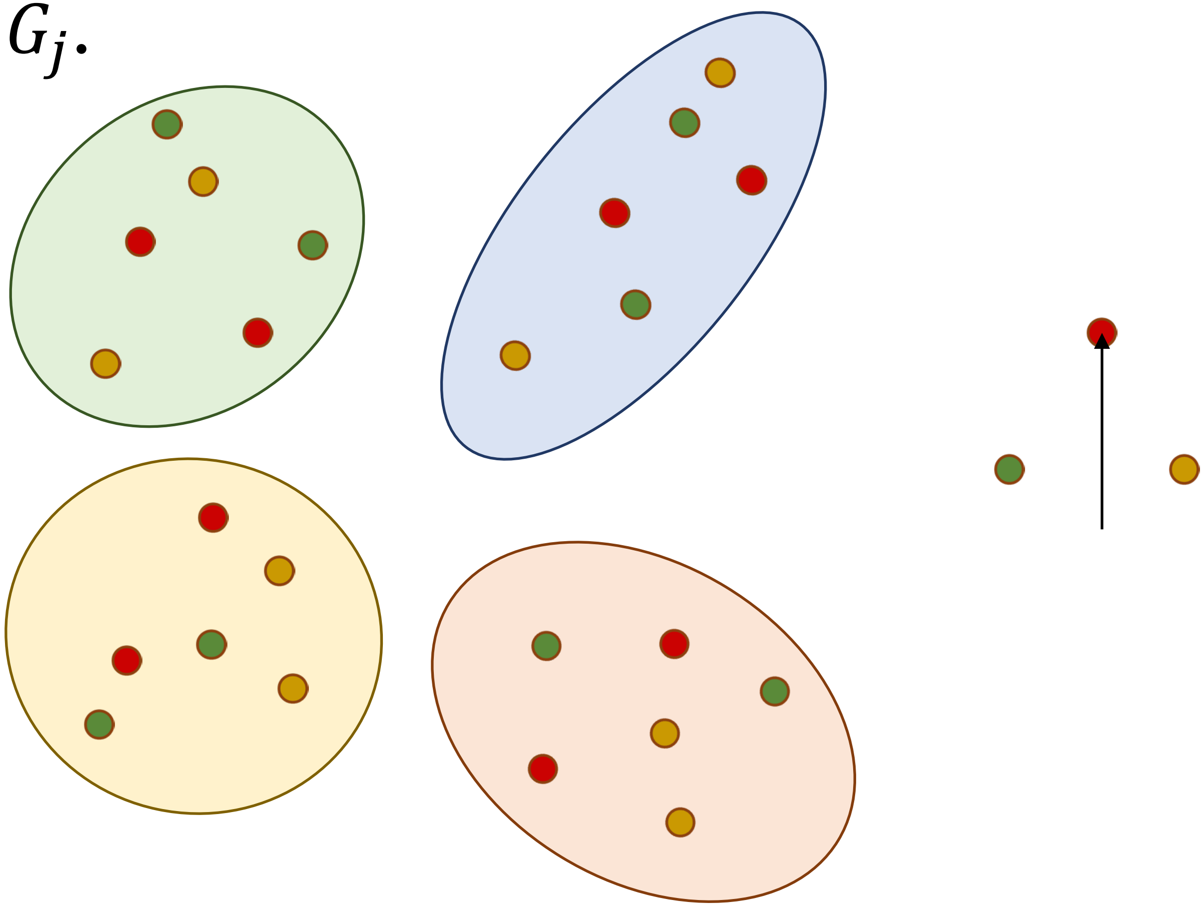
$$\ell \cdot \mathrm{cost}(Y, C_i^*, c_i^*) \leq OPT$$
$$\mathrm{fairlet}(F, Y) \leq 2OPT$$

Further the fairlet decomposition is consistent with the clustering: each F_i lies entirely in some C_j^*.

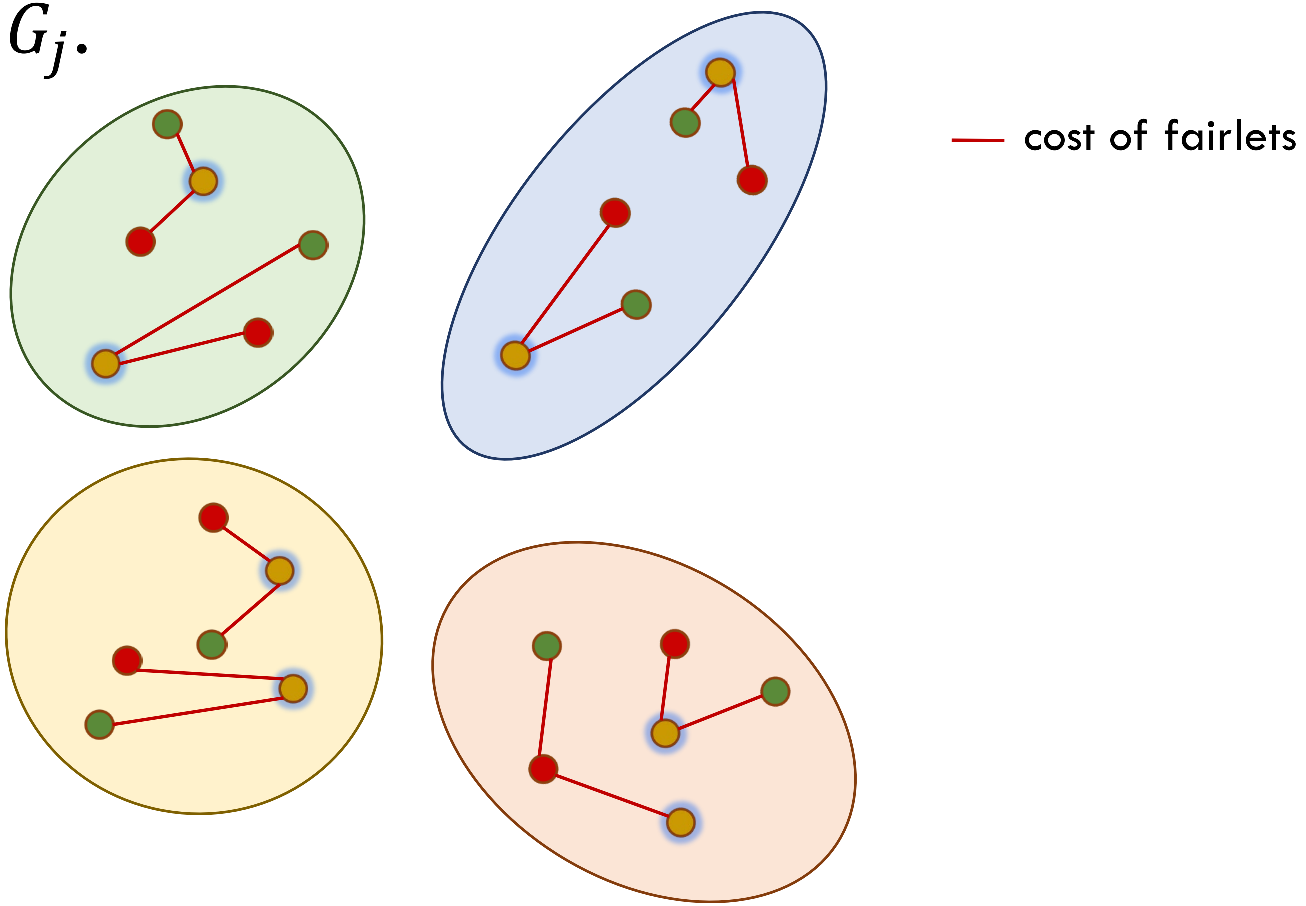
1. Each C_i^* has $|C_i^*|/\ell$ elements from each group. Partition it into $|C_i^*|/\ell$ fairlets arbitrarily.
2. Choose j uniformly at random from $\{1, \ldots, \ell\}$.
3. Let $Y = G_j$.



1. Each C_i^* has $\left|C_i^*\right|/\ell$ elements from each group. Partition it into $\left|C_i^*\right|/\ell$ fairlets arbitrarily.
2. Choose j uniformly at random from $\{1, \ldots, \ell\}$.
3. Let $Y = G_j$.



1. Each C_i^* has $|C_i^*|/\ell$ elements from each group. Partition it into $|C_i^*|/\ell$ fairlets arbitrarily.
2. Choose j uniformly at random from $\{1, \ldots, \ell\}$.
3. Let $Y = G_j$.



Proof of Lemma 2

Denote the center u is assigned to in the optimal solution by $c^*(u)$. Then

$$\mathbb{E}\mathrm{cost}(Y) = \frac{1}{\ell}\sum_j \mathrm{cost}(G_j) = \frac{1}{\ell}\sum_j \sum_{u \in G_j} d\big(u, c^*(u)\big) = \frac{OPT}{\ell}$$

⇒ For some j^* and $Y = E_{j^*}$:

$$\mathrm{cost}(Y) \le \frac{OPT}{\ell}$$

C_i^* is the optimal fair clustering

Proof of Lemma 2

Lemma 2: There exists a fairlet decomposition $F_1, \dots, F_k$ with representatives $Y = G_{j^*}$ for some j^* s.t.

$$\ell \cdot \mathrm{cost}(Y, C_i^*, c_i^*) \le OPT$$

$$\mathrm{fairlet}(F, Y) \le 2OPT$$

C_i^* is the optimal fair clustering

Proof of Lemma 2

Lemma 2: There exists a fairlet decomposition $F_1, \dots, F_k$ with representatives $Y = G_{j^*}$ for some j^* s.t.

$$\ell \cdot \mathrm{cost}(Y, C_i^*, c_i^*) \le OPT$$

$$\Rightarrow \quad \mathrm{fairlet}(F, Y) \le 2OPT$$

Proof of Lemma 2

$$
\begin{aligned}
\text{fairlet}(F, Y) &= \sum_{i=1}^{t} \sum_{u \in F_i} d(u, y_i) \\
&\leq \sum_{i=1}^{t} \sum_{u \in F_i} d\big(u, c^*(y_i)\big) + d\big(y_i, c^*(y_i)\big) \\
&= \sum_{i=1}^{t} \sum_{u \in F_i} d\big(u, c^*(u)\big) + d\big(y_i, c^*(y_i)\big) \\
&= OPT + \ell \cdot \text{cost}(Y) \leq 2\, OPT
\end{aligned}
$$

QED

Algorithm

Lemma 2: For some j^*, $Y = G_j$ and fairlet decomposition F:

$$\ell \cdot \mathrm{cost}\left(Y, C_i^*, c_i^*\right) \leq OPT$$

$$\mathrm{fairlet}(F, Y) \leq 2OPT$$

Algorithm Overview

- guess j^* and let $Y = G_{j^*}$
- let $C_1, \ldots, C_k$ be an approx. optimal clustering for Y
- let $\{F_i\}$ be an optimal choice of fairlets for Y
- extend the clustering to fairlets

Algorithm

Use an α approx. algorithm for standard k-medians

Lemma 2: For some j^*, $Y = G_j$ and fairlet decomposition F:

$$\ell \cdot \mathrm{cost}\left(Y, C_i^{ALG}, c_i^{ALG}\right) \leq \alpha \cdot OPT$$

$$\mathrm{fairlet}(F, Y) \leq 2OPT$$

Algorithm Overview

- guess j^* and let $Y = G_{j^*}$
- let $C_1, \dots, C_k$ be an approx. optimal clustering for Y
- let $\{F_i\}$ be an optimal choice of fairlets for Y
- extend the clustering to fairlets

Algorithm

Find optimal fairlets F^{ALG}

Lemma 2: For some j^*, $Y = G_j$ and fairlet decomposition F:

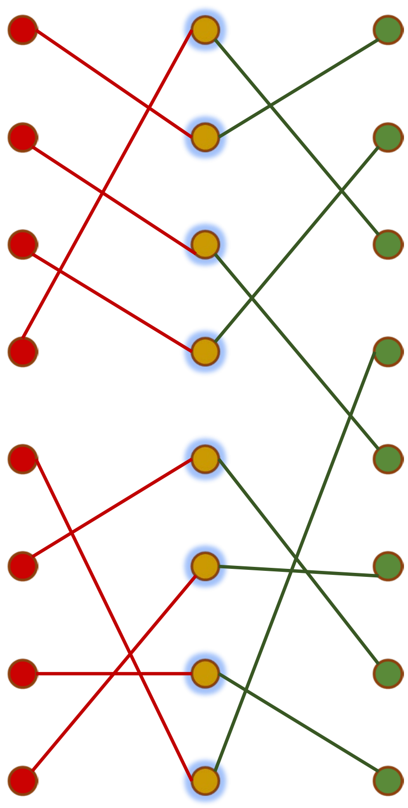
$$\ell \cdot \text{cost}\left(Y, C_i^{ALG}, c_i^{ALG}\right) \leq \alpha \cdot OPT$$

$$\text{fairlet}\left(F^{ALG}, Y\right) \leq 2OPT$$

Algorithm Overview

- guess j^* and let $Y = G_{j^*}$
- let $C_1, \ldots, C_k$ be an approx. optimal clustering for Y
- let $\{F_i\}$ be an optimal choice of fairlets for Y
- extend the clustering to fairlets

Finding fairlets: Min Cost Matching



Summary of the Algorithm

We have,

$$\ell \cdot \text{cost}\left(Y, C_i^{ALG}, c_i^{ALG}\right) \leq \alpha \cdot OPT$$

$$\text{fairlet}\left(F^{ALG}, Y\right) \leq 2OPT$$

Thus,

$$\text{cost}\left(X, C^{ALG}\right) \leq \ell \cdot \text{cost}(Y) + \text{fairlet}(F^{ALG}, Y)$$

$$\leq \alpha \cdot OPT + 2OPT = (\alpha + 2)OPT$$

QED

General Setting

[Bera, Chakrabarty, Flores, and Negahbani '19]
[Bercea, Groß, Khuller, Kumar, Rösner, Schmidt, and Schmidt '18]

Solve k-medians subject to general fairness constraints:

$$\alpha_j |C_i| \leq \left|C_i \cap G_j\right| \leq \beta_j |C_i|$$

No polynomial-time true approximation algorithms are known for arbitrary ℓ (which may depend on n).

Bera et al. & Bercea et al.:

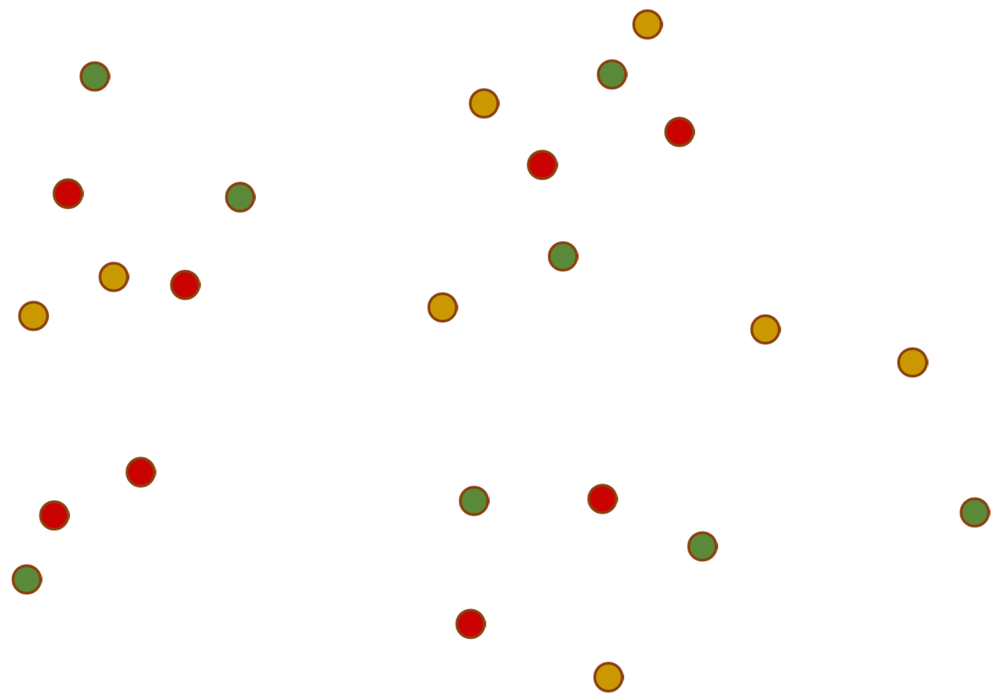
Find a solution that only slightly violates the fair representation constraints!

$$\alpha_j |C_i| - O(1) \leq \left|C_i \cap G_j\right| \leq \beta_j |C_i| + O(1)$$

Reduce the number of locations

Goal 1: Reduce the number of locations to k.

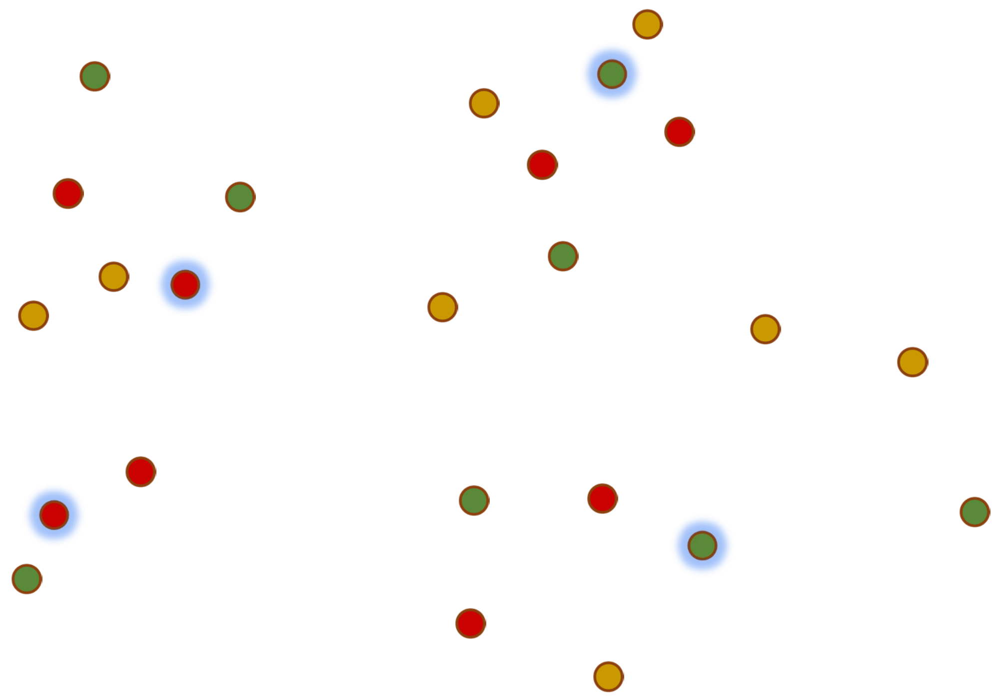
Solve standard k-medians and move each point to the closest center.



Reduce the number of locations

Goal 1: Reduce the number of locations to k.

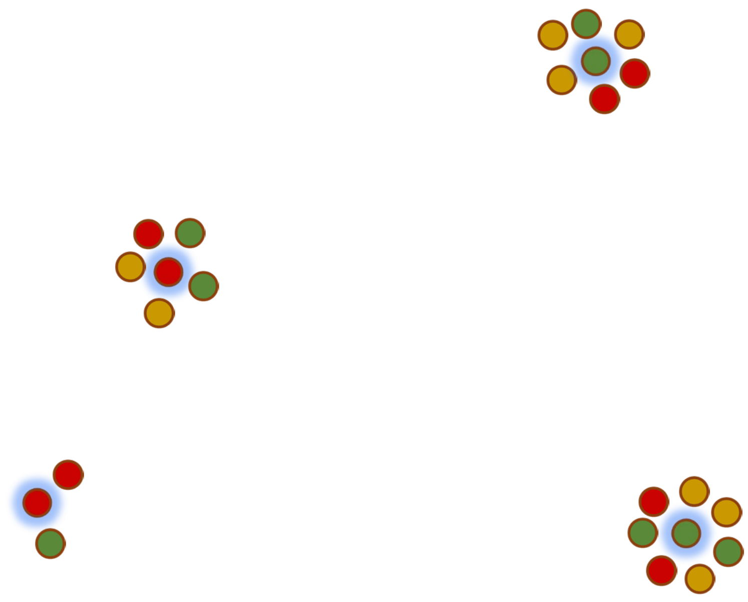
Solve standard k-medians and move each point to the closest center.



Reduce the number of locations

Goal 1: Reduce the number of locations to k.

Solve standard k-medians and move each point to the closest center.

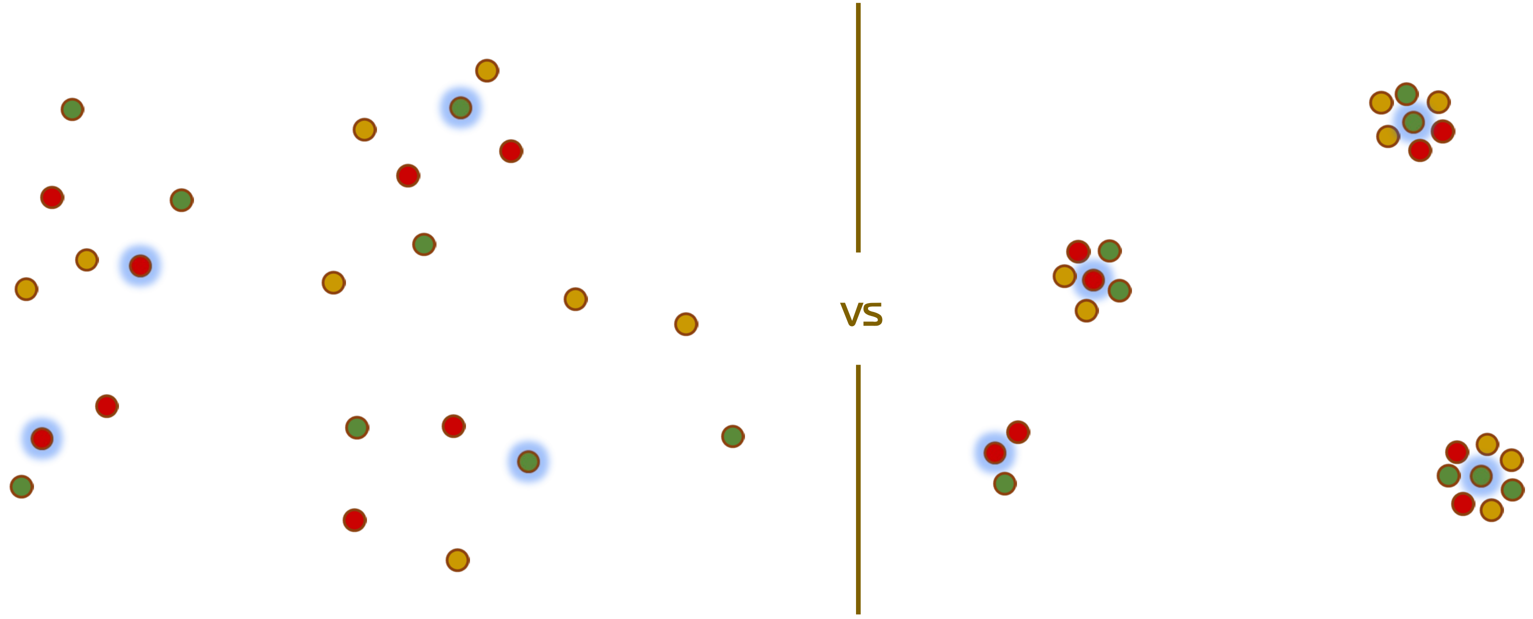


The current clustering is not fair!

Reduce the number of locations to k

The total distance travelled by all points is

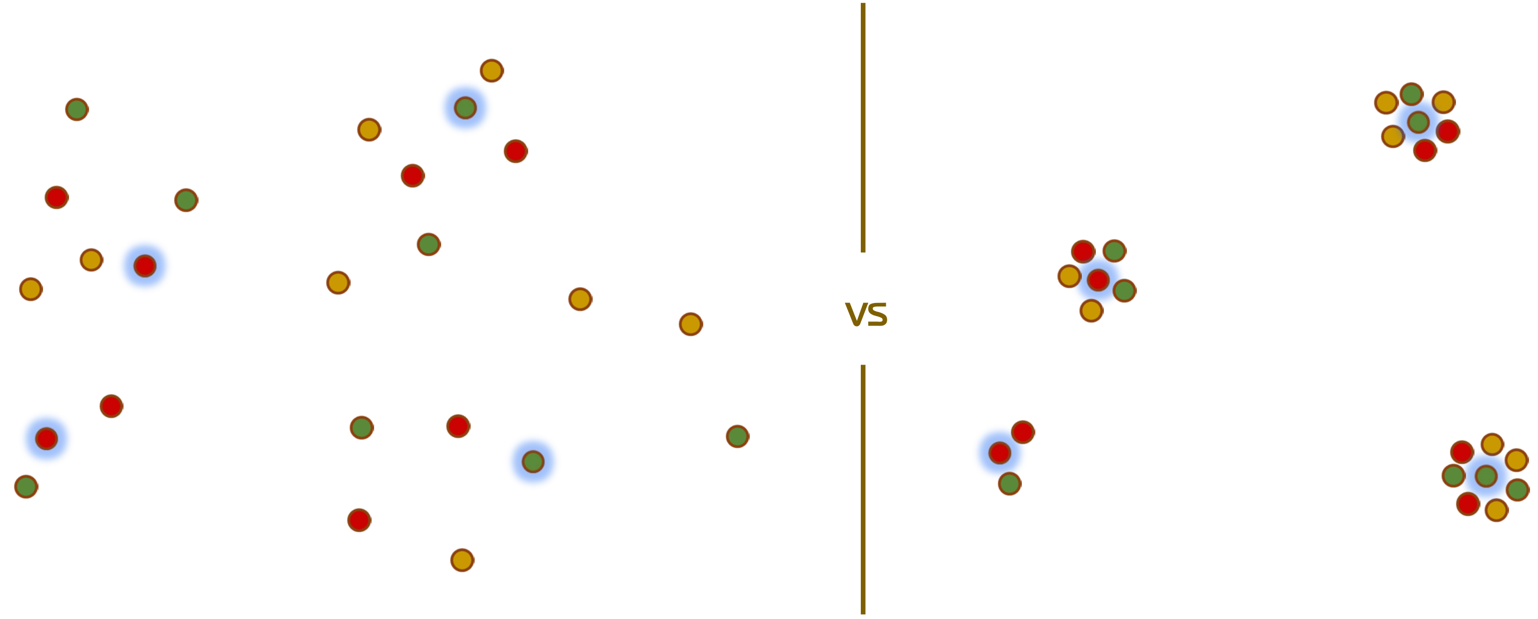
$$\alpha \cdot OPT_{\text{standard}} \leq \alpha \cdot OPT_{fair}$$



Reduce the number of locations to k

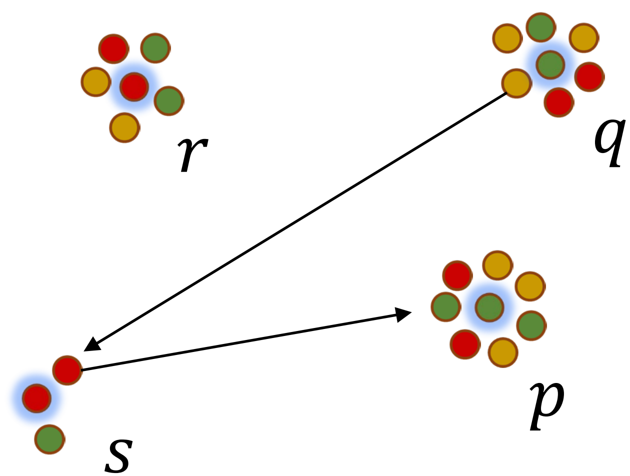
$\Rightarrow \left|\mathrm{cost}_{\mathrm{orig}}(C_1, \dots, C_k) - \mathrm{cost}_{\mathrm{new}}(C_1, \dots, C_k)\right| \leq O(OPT)$

$\Rightarrow$ A β approximation for the new instance gives an $O(\beta)$ approximation to the original one.



Reassignment Problem

Goal 2: Solve the obtained instance: reassign x_{abj} vertices of color j from center a to center b, so that the obtained clustering clustering is fair.



$x_{qs} = 1$

$x_{sp} = 1$

$x_{rr} = 2$

$\dots$

LP Relaxation

Let q_{aj} be the number of points of color j at center a.

minimize $\sum_{a,b\in[k],j\in[\ell]} d_{ab}x_{abj}$

$$
\begin{aligned}
&\textstyle\sum_b x_{abj} = q_{aj} && \forall a, j \\
&\alpha_j \textstyle\sum_{a,j'} x_{abj'} \le \sum_a x_{abj} \le \beta_j \sum_{a,j'} x_{abj'} && \forall b, j \\
&x_{abj} \ge 0 && \forall a, b, j
\end{aligned}
$$

Solve this LP and find a fractional solution x_{abj}.

LP Refractional → Integrallaxation

☹ Solution x_{abj} is not necessarily integral.

Goal: Find a "similar" integral solution f_{abj}.

Want: solution f assigns to each center b approximately the same number of points of each color j as x does.

LP Refractional → Integrallaxation

☹ Solution x_{abj} is not necessarily integral.

Goal: Find a “similar” integral solution f_{abj}.

Want: solution f assigns to each center b approximately the same number of points of each color j as x does.

According to x:

$\sum_{a,j'} x_{abj'}$	centers are assigned to center b
$\sum_{a} x_{abj}$	centers of color j are assigned to b

These numbers are not necessarily integers.

Fractional → Integral

According to the LP solution:

$\sum_{a,j'} x_{abj'}$ centers are assigned to center b

$\sum_a x_{abj}$ centers of color j are assigned to b

Let

$$l_b = \left\lfloor \sum_{a,j'} x_{abj'} \right\rfloor \text{ and } u_b = \left\lceil \sum_{a,j'} x_{abj'} \right\rceil$$

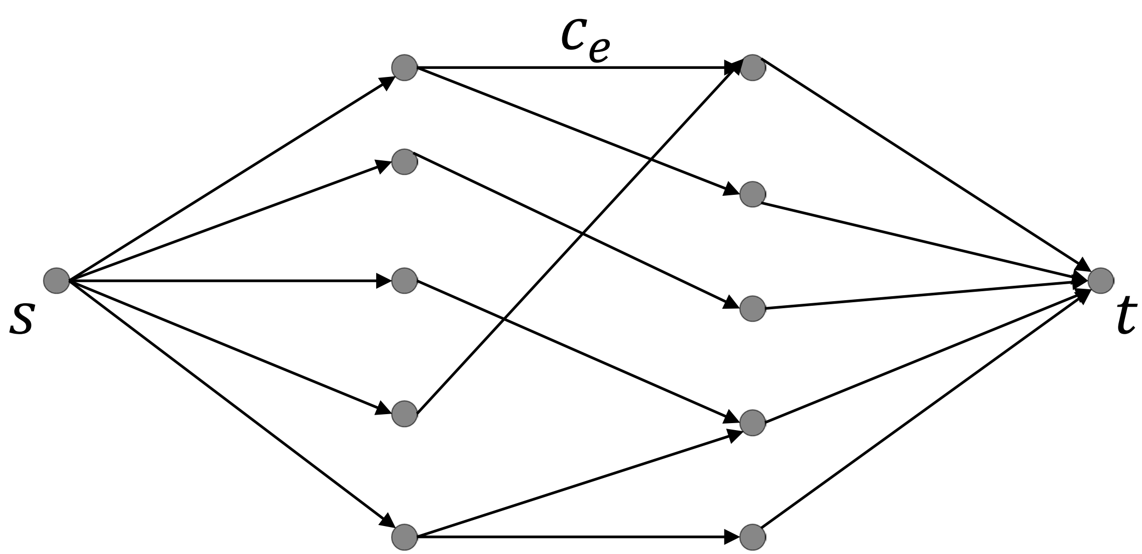
$$l_{bj} = \left\lfloor \sum_a x_{abj} \right\rfloor \text{ and } u_{bj} = \left\lceil \sum_a x_{abj} \right\rceil$$

Goal: find a solution with cluster C_b of size $|C_b| \in [l_b, u_b]$ and $|C_b \cap G_j| \in [l_{bj}, u_{bj}]$.

Minimum Cost s-t Flow

Given: an s-t flow network with edge capacities c_e, costs d_e and a parameter F.

Goal: Send F units of flows from s to t subject to capacity constraints so as to minimize $\sum_e d_e f_e$.

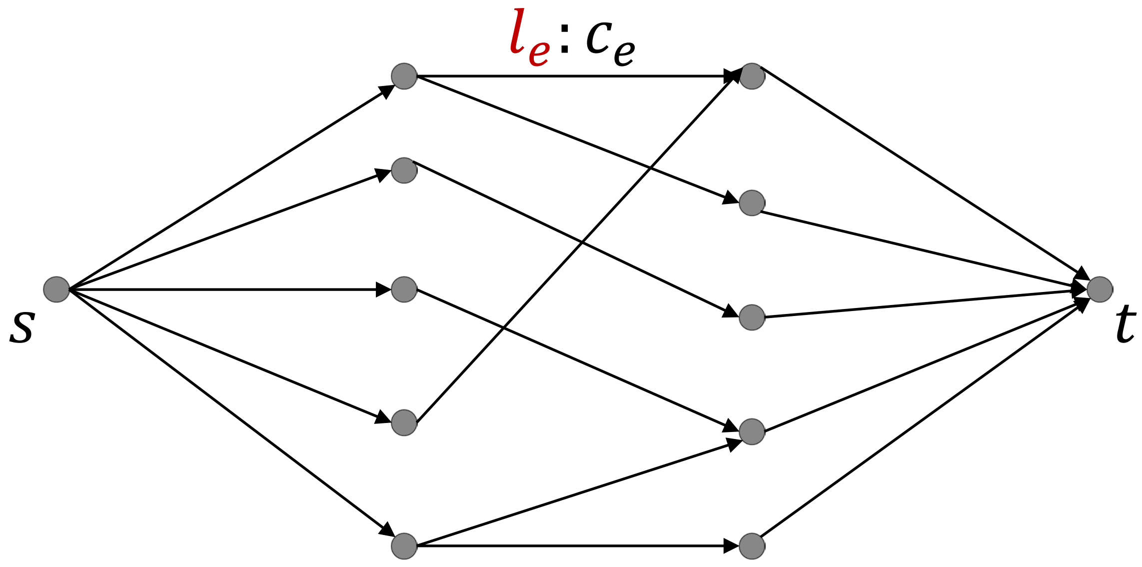


Note: There exists an integral optimal flow as long as all c_e and F are integral! It can be found in poly time.

Minimum Cost s-t Flow

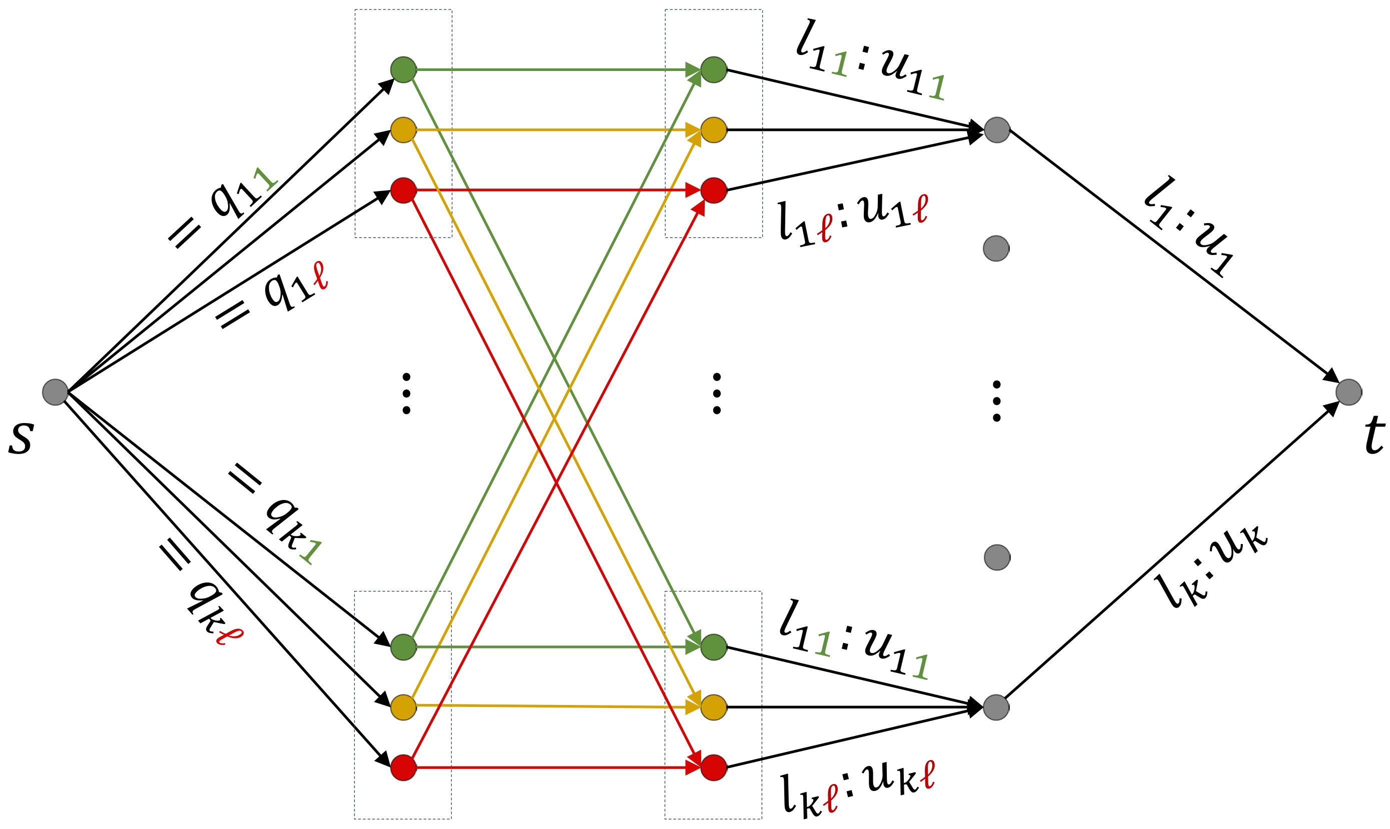
Given: an s-t flow network with edge capacities c_e and lower bounds l_e, and costs d_e.

Goal: Send flow from s to t subject to capacity and lower bound constraints so as to minimize $\sum_e d_e f_e$.

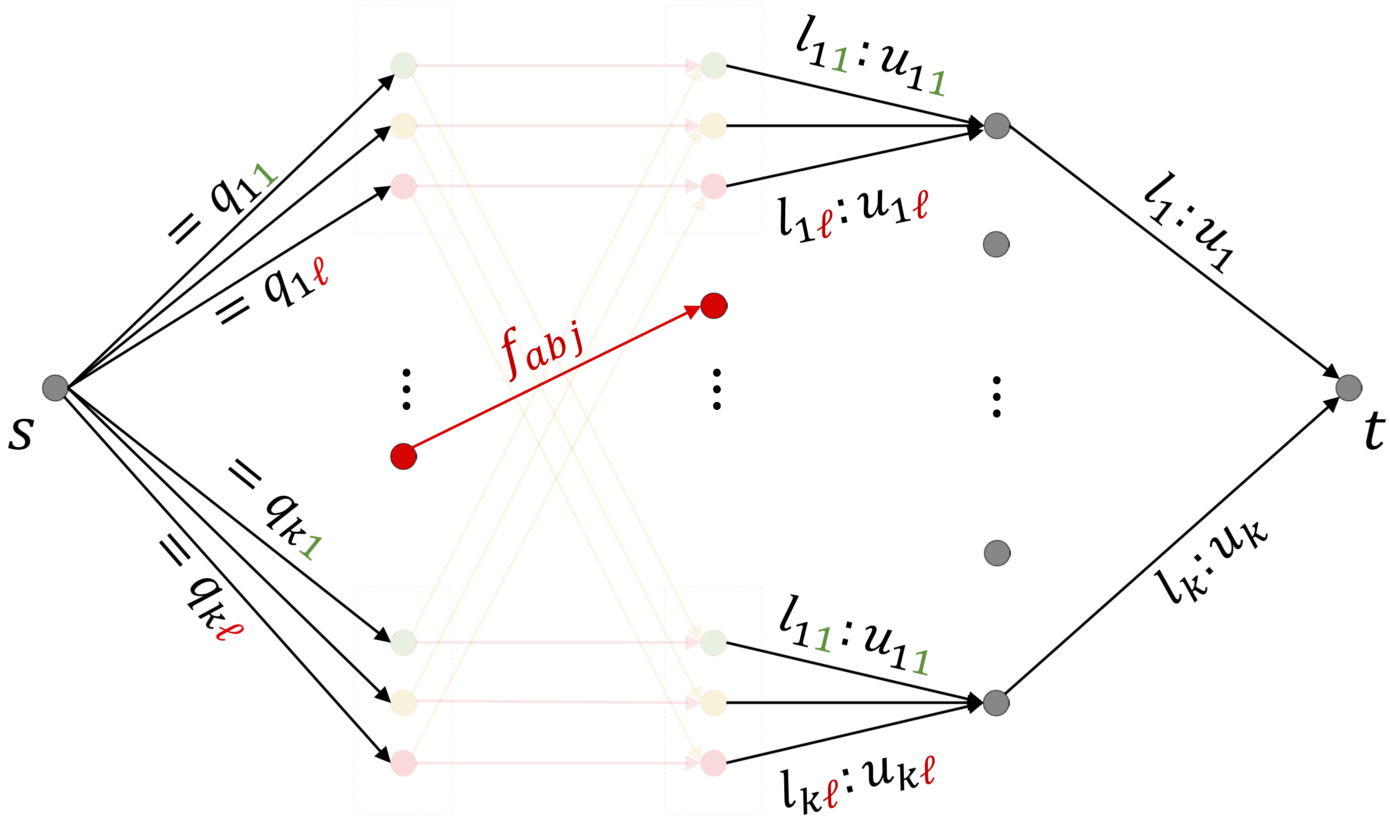


Note: There exists an integral optimal flow as long as all c_e and l_e are integral! It can be found in poly time.

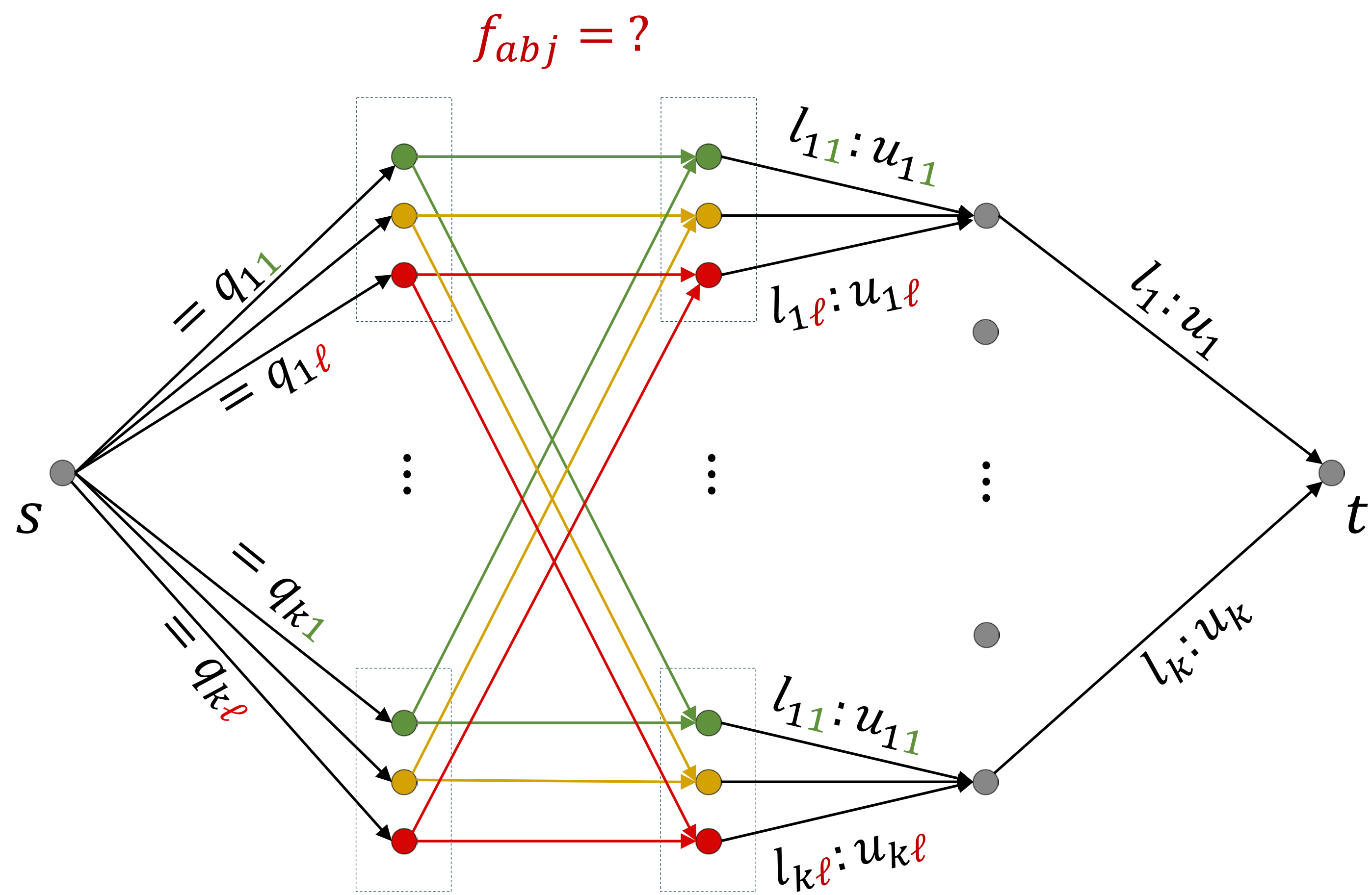
Minimum Cost s-t Flow



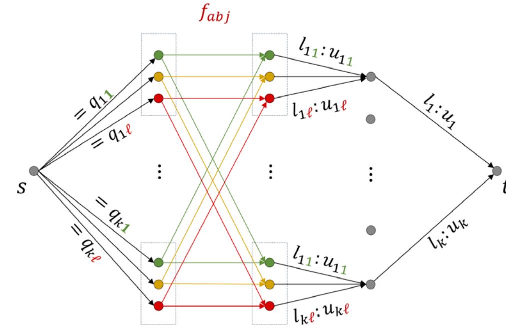
Minimum Cost s-t Flow



Q: Is there a feasible flow?



Fractional → Integral

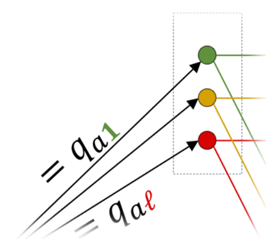


$l_b = \lfloor \sum_{a,j'} x_{abj'} \rfloor$ and $u_b = \lceil \sum_{a,j'} x_{abj'} \rceil$

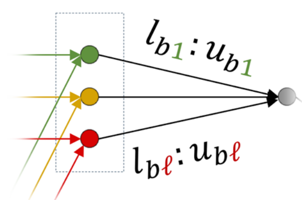
$l_{bj} = \lfloor \sum_a x_{abj} \rfloor$ and $u_{bj} = \lceil \sum_a x_{abj} \rceil$

We get integer f_{abj} s.t.

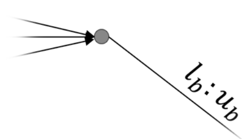
$$\sum_b f_{abj} = q_{aj}$$



$$\sum_a f_{abj} \in [l_{bj}, u_{bj}]$$



$$\sum_{aj'} f_{abj'} \in [l_b, u_b]$$



Summary

We reassign f_{abj} points of color j from a to b.

- The cost of this solution is at most the cost of the LP

$$O(\alpha \cdot OPT)$$

- Fairness constraints:

$$\sum_a f_{abj} \geq l_{bj} = \left\lfloor \sum_a x_{abj} \right\rfloor > \sum_a x_{abj} - 1 \geq \alpha_j \sum_{a,j'} x_{abj'} - 1$$

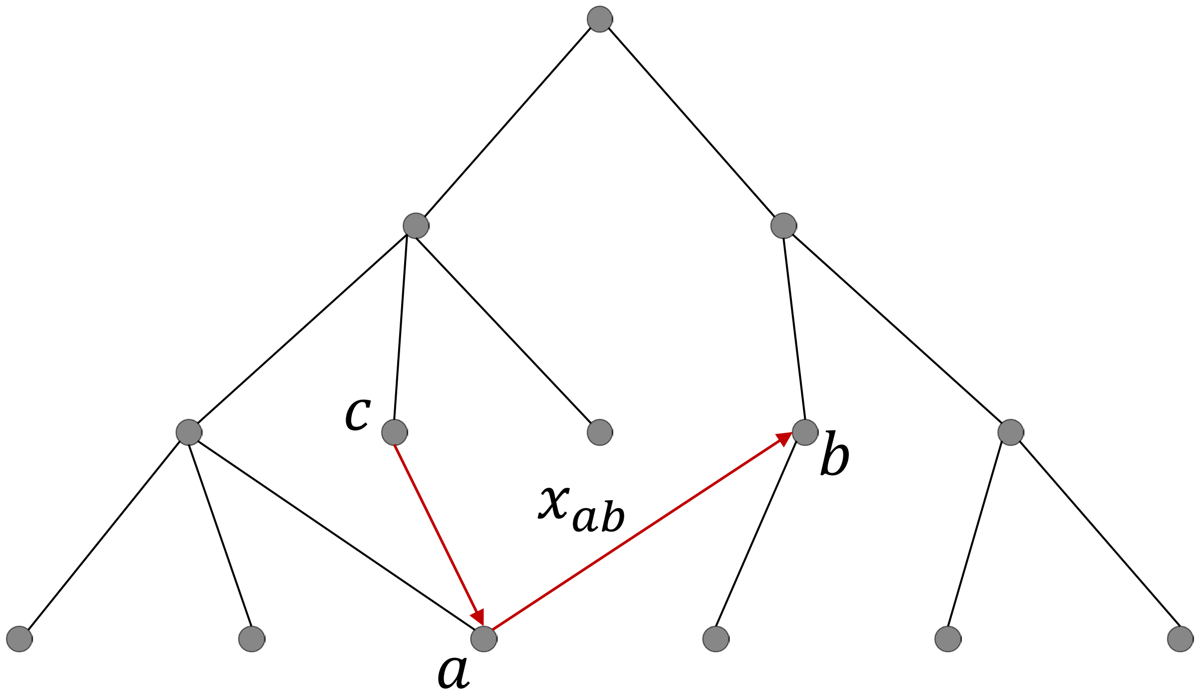
$$\geq \alpha_j l_b - 1 \geq \alpha_j (u_b - 1) - 1 \geq \alpha_j \sum_{aj'} f_{abj'} - 2$$

Solving Reassignment Exactly for fixed ℓ [Dai, M, Vakilian '22]

Assume that the number of groups ℓ is a small integer.

- Embed metric d_{uv} on $\{c_1, \dots, c_k\}$ into a distribution of dominating trees with distortion $O(\log k)$.
- Sample tree T from the distribution.
- Reassignment problem. Denote

$$x_{ab} = \begin{pmatrix} x_{ab1} \\ x_{ab2} \\ \vdots \\ x_{abj} \end{pmatrix} \quad \text{and} \quad q_a = \begin{pmatrix} q_{a1} \\ q_{a2} \\ \vdots \\ q_{aj} \end{pmatrix}$$



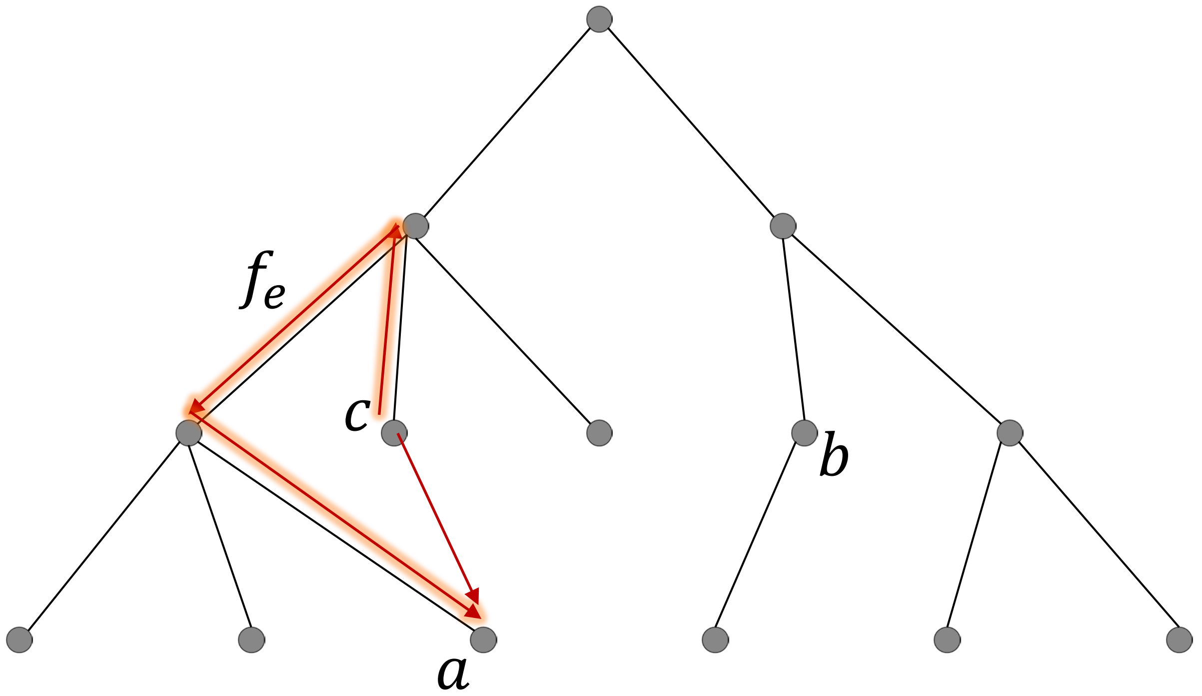
Total assignment to vertex a is

$$A_a = q_a + \sum_{c \neq a} x_{ca} - \sum_{b \neq a} x_{ab}$$

Fairness constraint:

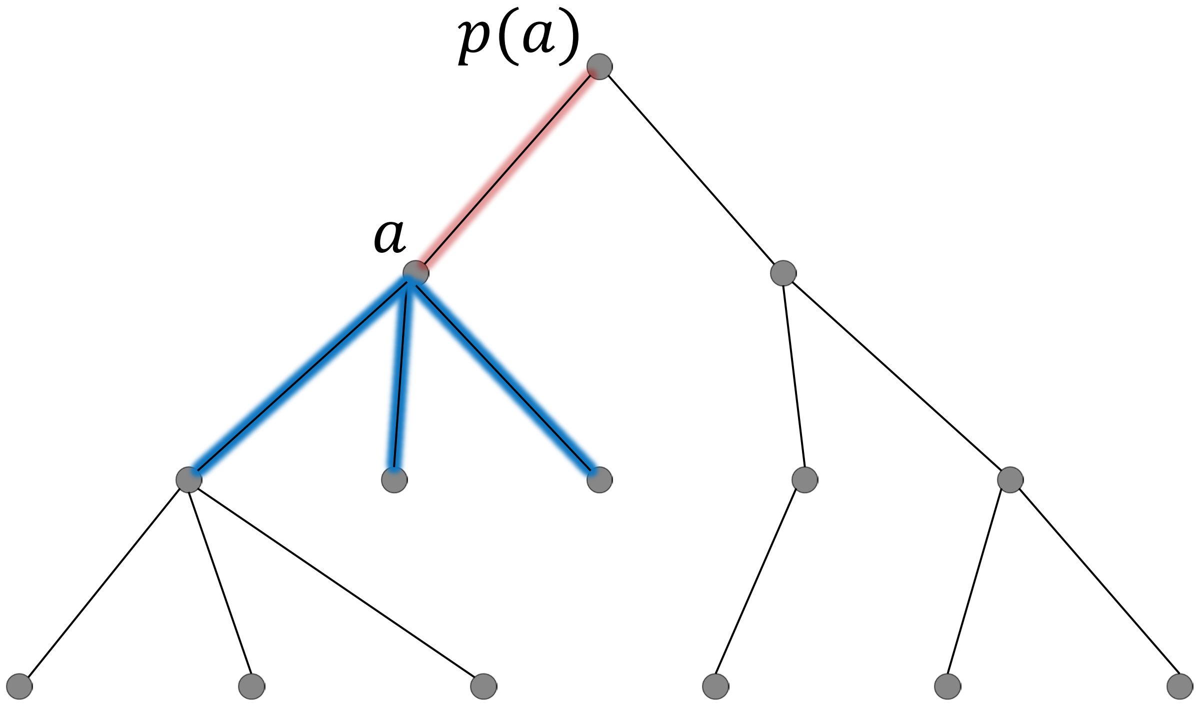
$$\alpha_j \|A_a\|_1 \leq A_{aj} \leq \beta_j \|A_a\|_1$$

Cost: $\sum_{ab} d_T(a, b) \cdot \|x_{ab}\|_1$



Reroute the flow along the tree edges. Let f_{ej} be the net amount of flow of type j going down along edge e. f_{ej} may be positive or negative.

$$f_e = \begin{pmatrix} f_{e1} \\ \cdots \\ f_{e\ell} \end{pmatrix}$$



Total assignment to vertex a is

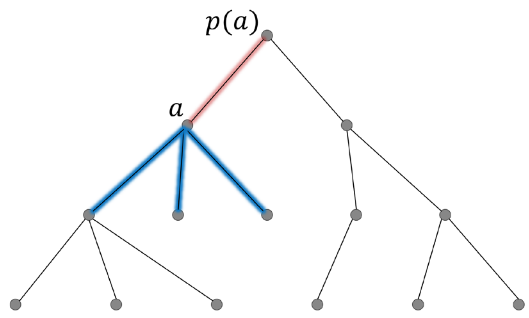
$$A_a = q_a + f_{p(a)a} - \sum_{b:p(b)=a} f_{ab}$$

Fairness constraint:

$$\alpha_j \|A_a\|_1 \le A_{aj} \le \beta_j \|A_a\|_1$$

Cost: $\sum_{(a,b)\in T} d(a,b) \cdot \|f_{ab}\|_1$

Dynamic Programming



Find flow $f_e \in \{-n, \dots, n\}^{\ell}$ so that

$$\alpha_j \|A_a\|_1 \le A_{aj} \le \beta_j \|A_a\|_1$$

for every a, where

$$A_a = q_a + f_{p(a)a} - \sum_{b:p(b)=a} f_{ab}$$

so as to minimize the total cost

$$\sum_{(a,b)\in T} d(a,b) \cdot \|f_{ab}\|_1$$

This can be done in time $n^{O(\ell)}$.