Individual Fairness in Clustering

High-Level Motivation

- **Demographic Fairness:** Treat each group of points fairly, with respect to how other groups are being treated or with respect to the specific needs of the group at hand.
- **Individual Fairness:** Treat each individual point fairly, with respect to how other points are being treated or with respect to its specific needs.
- Does demographic fairness imply individual fairness?
  - View each point as a singleton group.
  - The concepts of group fairness become vague or ill-defined in this case:
    - Balance: Leads to a single cluster solution
    - Proportionality: Each point is entitled to each its own cluster?
    - Socially fair k-clustering: Reduces to k-center
- Demographic fairness cannot adequately capture any individual needs of points.

The Seminal Work of Dwork et al.

- A very important work in the area of Individual Fairness
- Dwork et al. ("Fairness Through Awareness" – ITCS 2012) introduced a ground breaking concept of individual fairness in the context of classification.

Similar individuals should be treated similarly

- It will help us in our taxonomy of individually fair notions for clustering
  1) Definitions that follow the Dwork et al. paradigm
  2) Definitions that diverge from it

Individually-Fair Clustering Models that Follow the Dwork et al. Paradigm

The Dwork et al. Paradigm in Clustering

Similar individuals should be treated similarly

- Two questions that need to be answered:
  1) How can we define similarity in the context of clustering?
  2) What constitutes similar treatment in a clustering setting?
- The first question is not really important.
- The second question is of much more significance.

Similar treatment in terms of same cluster placement

- **Motivational Example:**
  - Suppose a company wants to cluster its employees into k groups
  - People in the first cluster will receive the highest amount of raise, the people in the second cluster the second highest raise, and so on.
  - Suppose that employee X is very similar to employee Y.
  - If Y is placed in a cluster that receives a better amount of raise, then X **would arguably feel unfairly treated.**

- In such cases, similar points should be placed in the same cluster

Probabilistic Pairwise Fairness – Definition of Similarity

- Introduced by Brubach et al. ("A Pairwise Fair and Community-preserving Approach to k-Center Clustering" – ICML 2020)
- Definition of similarity:
  - For every pair of points $j, j' \in \mathcal{C}$ we are given a value $\psi_{j,j'} \in [0,1]$ indicating their true similarity.
  - The smaller $\psi_{j,j'}$ is the more similar the two points.
- The values ψ can be different from the metric d:
  1) Encoding of redundant features in d
  2) ψ can be the similarity as perceived by the individuals

Probabilistic Pairwise Fairness – Definition of Similar Treatment

- How can we mitigate unfair behavior?
- Avoid situations where two similar points are deterministically separated

Randomization can imply fairness

- Seek a randomized solution that separates j, j' with probability at most $\psi_{j,j'}$
  - Choose S with $|S| \leq k$
  - Construct efficiently sampleable distribution $\mathcal{D}$ over assignments $\varphi: \mathcal{C} \rightarrow S$ such that

$$\Pr_{\varphi \sim \mathcal{D}}[\varphi(j) \neq \varphi(j')] \leq \psi_{j,j'}$$

  - Minimize some metric related objective

Probabilistic Pairwise Fairness - Results

- Brubach et al. (“A Pairwise Fair and Community-preserving Approach to k-Center Clustering” – ICML 2020) introduced the problem and gave a $logk$-approximation algorithm for the k-center objective.
  - The algorithm works when $\psi_{j,j'} = \{\frac{d(j,j')}{R}, 1\}$, for some $R > 0$.
  - Very efficient algorithm
  - Bounded PoF
- Brubach et al. (“Fairness, Semi-Supervised Learning, and More: A General Framework for Clustering with Stochastic Pairwise Constraints” – AAAI 2021) gave constant factor approximations for all k-center, k-median and k-means
  - The values $\psi_{j,j'}$ are arbitrary
  - Not that efficient – LP based

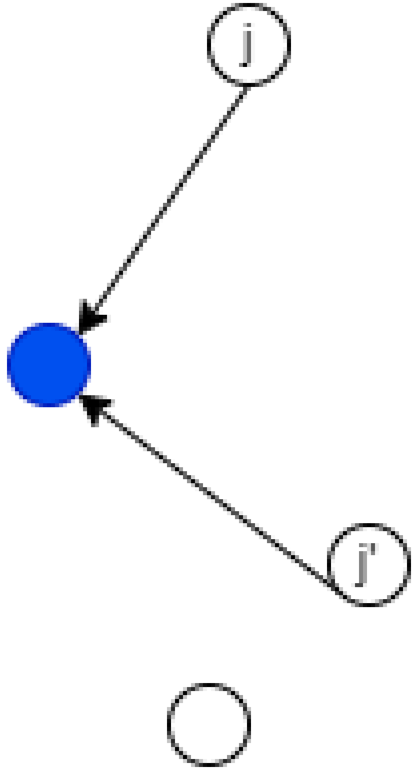
Distributional Individual Fairness

- Introduced by Anderson et al. ("Distributional Individual Fairness in Clustering" – Arxiv 2020).
- Similarity defined exactly as in Brubach et al. That is with values $\psi_{j,j'}$
- Pick $S \subseteq \mathcal{C}$ with $|S| \leq k$
- For each $j \in \mathcal{C}$ find distribution φ_j over S
- Fairness constraint:
  - Metric D measuring statistical proximity
  - $D(\varphi_j, \varphi_{j'}) \leq \psi_{j,j'}$
- Difference with the model of Brubach et al.
  - Brubach et al. return an actual assignment $\varphi: \mathcal{C} \to S$
  - Brubach et al. upper bound the separation probability
    - Example: For j, j' both φ_j and $\varphi_{j'}$ are the uniform distribution over S
- Anderson et al. give constant factor approximation algorithms for all k-center, k-median and k-means

Similar Treatment is Terms of the Assignment Distance

- In many applications the quantity $d(j,\varphi(j))$ (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.
- Suppose j' is similar to j and $d(j',\varphi(j')) \ll d(j,\varphi(j))$.

j is justified to feel unfairly treated



Motivational Example

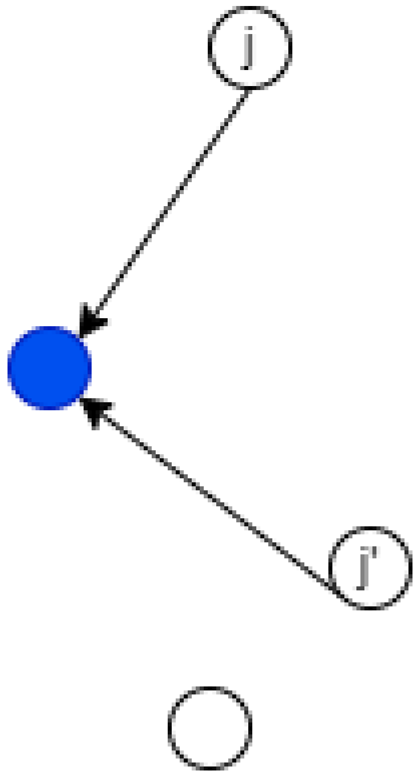
- The points of $\mathcal{C}$ correspond to users of an e-commerce site.
- $d(j, j')$ measures how similar the profiles of j and j' are.
- The website wants to choose k representative users $S \subseteq \mathcal{C}$ (according to some objective function) and construct an assignment $\varphi: \mathcal{C} \rightarrow S$.
- User j will receive recommendations based on $\varphi(j)$'s profile.
- The smaller $d\big(j, \varphi(j)\big)$ is the more relevant the recommendations j receives.
- If j considers j' as similar to itself, then it perceives $d\big(j', \varphi(j')\big) \ll d\big(j, \varphi(j)\big)$ as unfair treatment.

α-Equitable k-Center

- Introduced by Chakrabarti et al. ("A New Notion of Individually Fair Clustering: α-Equitable k-Center" – AISTATS 2022)
- Every point j has a set of other points $\mathcal{S}_j \subseteq \mathcal{C}$ that it perceives as similar to itself
  - This is how similarity is modeled in this work
  - Has advantages over the modeling with the ψ values: more easily constructable
- We are also given a value $\alpha \geq 1$.
- Ask for $S \subseteq \mathcal{C}$ ($|S| \leq k$) and assignment $\varphi: \mathcal{C} \rightarrow S$ that minimize the k-center objective $\max_{j \in \mathcal{C}} d(j, \varphi(j))$.
- **Fairness Constraint:** For every $j \in \mathcal{C}$ and $j' \in \mathcal{S}_j$ ensure that $d\big(j, \varphi(j)\big) \leq \alpha \cdot d\big(j', \varphi(j')\big)$
  - The smaller α is the smaller $\frac{d(j,\varphi(j))}{d(j',\varphi(j'))}$ remains

The parameter α

- The smaller α is the smaller $\frac{d(j,\varphi(j))}{d(j',\varphi(j'))}$ remains.
  - α = 4
  - α = 1
- A value of α close to 1 would give the most equitable/fair solution
- For what values of α is the problem well-defined?
  - For $a < 2$ there exist instances that admit no feasible solution
  - For $a \geq 2$ we can always find a feasible solution



The results of Chakrabarti et al.

- A very efficient algorithms that returns a solution of cost $5(R^* + R_m)$
  - R^* is the value of the optimal solution
  - $R_m = \max_{j \in \mathcal{C}, j' \in \mathcal{S}_j} d(j, j')$

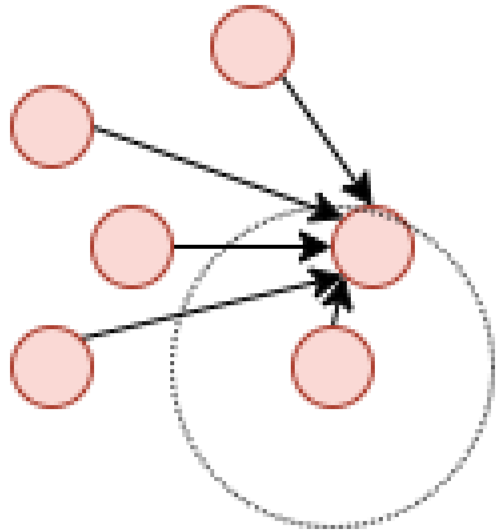
- When d is a good estimate of similarity: $R_m = O(R^*)$

- Under some mild conditions on the sets $\mathcal{S}_j$ the algorithm has bounded PoF

Notions of Individual Fairness in Clustering that do not follow the Dwork et al. paradigm

A Center in my Neighborhood

- Suppose we want to solve a classical k-clustering problem on a set of points $\mathcal{C}$
  - Find $S \subseteq \mathcal{C}$ ($|S| \leq k$) and assignment $\varphi : \mathcal{C} \rightarrow S$ that $\sum_{j \in \mathcal{C}} d(j, \varphi(j))^p$ is minimized
- Even though the global objective function might be minimized, individual points may have different requirement in terms of $d(j, \varphi(j))$
  - Recall the vaccine site allocation example.
- Each j has a value r_j, and we should make sure that $d(j, \varphi(j)) \leq r_j$



Results

- Jung et al. ("A Center in Your Neighborhood: Fairness in Facility Location" – FORC 2020) introduced the problem
  - Important result: Even finding a feasible solution to the problem is NP-hard.

- Goal: Find (α, β)-bicriteria algorithms:
  - $\sum_{j \in C} d\big(j, \varphi(j)\big)^p \leq \alpha \cdot \mathrm{OPT}$
  - $d(j, \varphi(j)) \leq \beta \cdot r_j$ for every j

- A series of papers gave increasingly better results:
  1) Mahabadi and Vakilian ("Individual Fairness for k-Clustering"- ICML 2020). $(O(p), 7)$-bicriteria
  2) Chakrabarty and Negahbani ("Better Algorithms for Individually Fair k-Clustering" – NeurIPS 2021) $(2^{1+\frac{2}{p}}, 8)$-bicriteria
  3) Vakilian and Yalçıner ("Improved Approximation Algorithms for Individually Fair Clustering" – AISTATS 2022) $(16^p, 3)$-bicriteria

Individual Fairness in Clustering with Outliers

- Pick $S \subseteq \mathcal{C}$ with $|S| \leq k$
- Pick $\mathcal{A} \subseteq \mathcal{C}$ with $|\mathcal{A}| \geq m$ (points to be clustered)
- **Being an outlier is disadvantageous!!!**
- We have seen how to protect against demographic bias
- What can be interpreted as bias against individuals?

Deterministically be chosen as an outlier in every computed solution

Randomization saves the day: A lottery model for individually fair clustering with outliers

- For each $j \in \mathcal{C}$ we are given a value $p_j \in [0,1]$
- We want a distribution $\mathcal{D}$ over solutions $(S, \mathcal{A})$ such that:
  1) For every $(S, \mathcal{A})$ drawn from $\mathcal{D}$ we have $|S| \leq k$ and $|\mathcal{A}| \geq m$.
  2) $\Pr_{(S,\mathcal{A})\sim\mathcal{D}}[j \in \mathcal{A}] \geq p_j$ for every $j \in \mathcal{C}$
  3) Some objective is minimized
- We avoid scenarios where certain points are deterministically chosen as outliers
- Through the values p_j we can capture a plethora of fairness concepts:
  - Equitable treatment: p_j is the same for all points
  - Preferential treatment: Points in greater need of service get a higher p_j value

Results

- The problem has only been studied under the k-center objective.

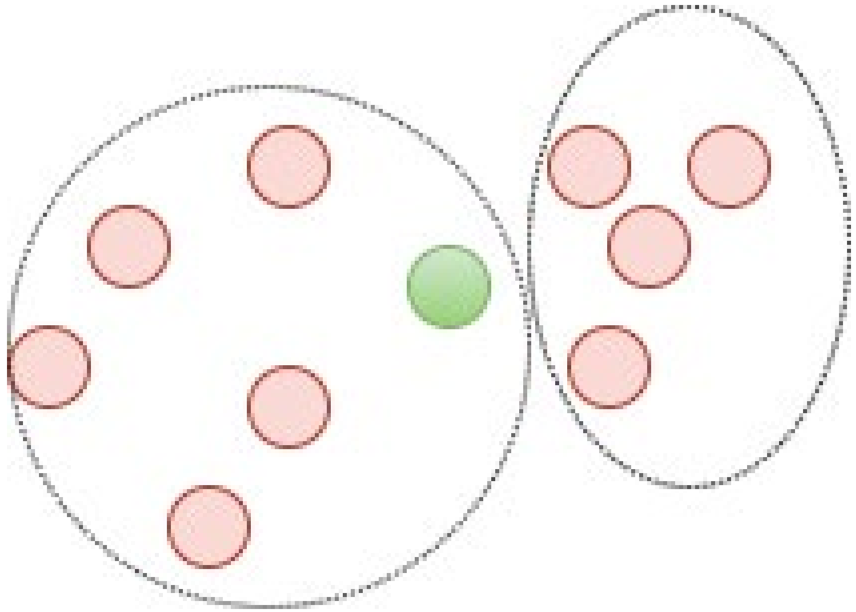
- It was introduced by Harris et al. (“A Lottery Model for Center-Type Problems With Outliers” – APPROX-RANDOM 2017)

- Harris et al. gave a pseudo 2-approximation algorithm.
  - In every solution drawn from $\mathcal{D}$ the coverage guarantee is $(1-\varepsilon)m$
  - $\Pr_{(S,\mathcal{A})\sim\mathcal{D}}[j \in \mathcal{A}] \geq (1-\varepsilon)p_j$

- Anegg et al. (“A Technique for Obtaining True Approximations for k-Center with Covering Constraints” – IPCO 2020) gave a true 4-approximation algorithm.

Fairness based on average distance to the points in your cluster

- **Motivational Example:**
  - Suppose a company wants to cluster its employees into k groups, based on their performance rating for some specific year.
  - Let's assume that people in the first cluster will receive the highest amount of raise, the people in the second cluster the second highest raise, and so on.
  - Consider some employee X placed in some cluster C. Let C_X be the average distance of X to the rest of the points in C.
  - If there exists cluster W, with W_X be the average distance of X to the of the points in W, such that $W_X \leq C_X$, **then X would arguably feel unfairly treated**



Formal Definition and Results

- Given a set of points $\mathcal{C}$, partition it into k sets $\mathcal{C}_1, \ldots, \mathcal{C}_k$ such that:
  - For every $i \in [k]$ and each j $\in \mathcal{C}_i$, $\frac{1}{|\mathcal{C}_i|-1}\sum_{j' \in \mathcal{C}_i} d(j,j') \leq \frac{1}{|\mathcal{C}_{i'}|}\sum_{j' \in \mathcal{C}_{i'}} d(j,j')$ for all $i' \neq i$

- The problem was introduced by Kleindessner et al. ("A Notion of Individual Fairness for Clustering" – Arxiv 2020).

- Main result: For $k \geq 2$, it is NP-hard to decide if such a clustering exists

- When the metric space is the Euclidean line, the problem can be solved efficiently.