

FairDen: Fair Density-Based Clustering

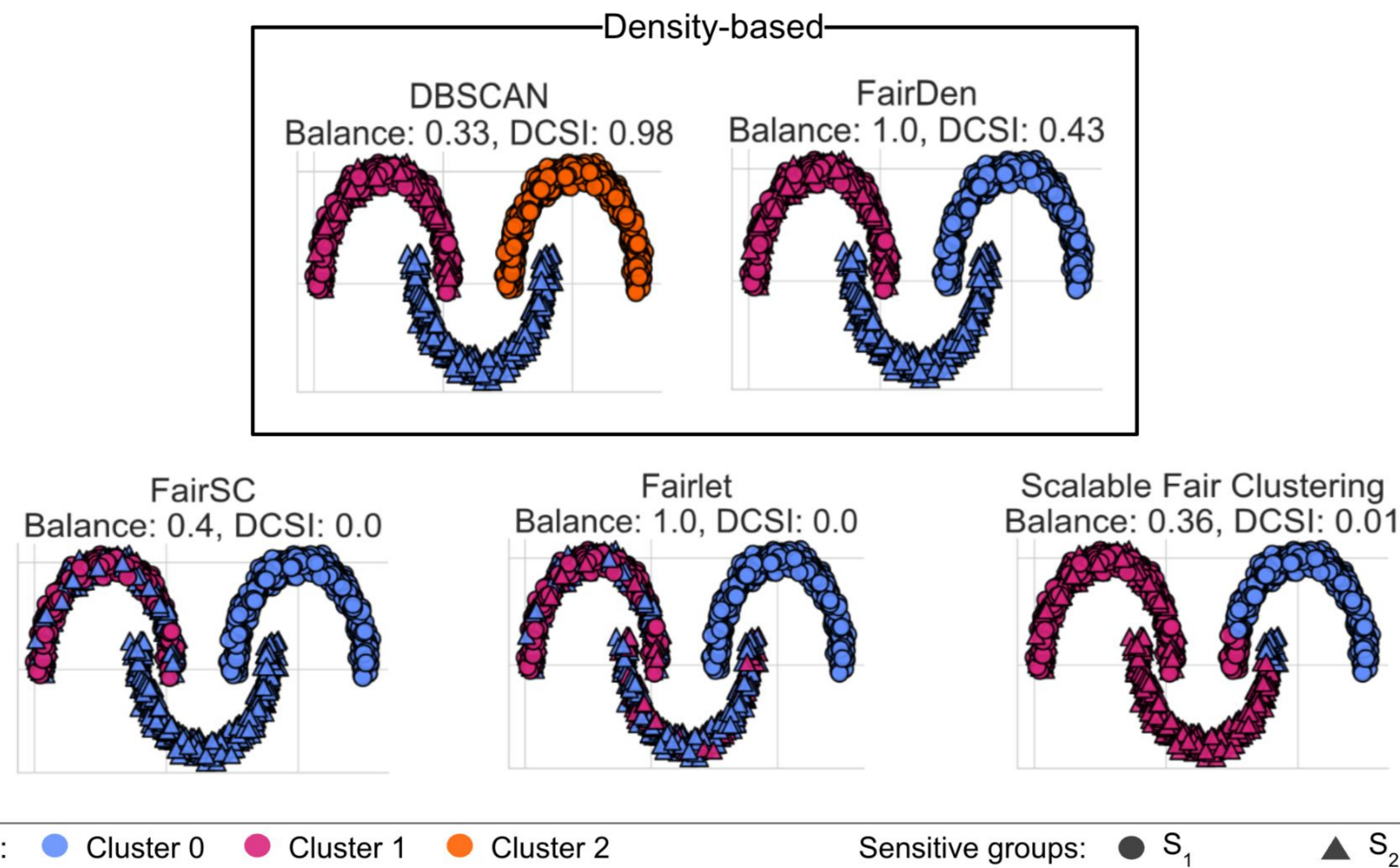
Lena Krieger^{*1}, Anna Beer^{*2}, Pernille Matthews³, Anneka Myrup Thieson³, Ira Assent^{1,3}

¹IAS-8, Forschungszentrum Jülich, Germany ²Institute for Computer Science, LMU Munich, Germany

³Faculty of Computer Science, University of Vienna, Austria ⁴Department of Computer Science, Aarhus University, Denmark

*Equal Contribution.

Fairness in Density-Based Clustering



For a binary sensitive attribute indicated by the shape, a **balanced** density-based cluster has the **same ratio** of sensitive groups (circles and triangles) as in the overall dataset. The three equally sized moons are density-based clusters that have different ratios: the first moon has 50% triangles and 50% circles, the second moon has only triangles, the third moon has only circles. We show the balance and DCSI values (higher is better), for five clustering methods: FairDen achieves optimal balance and a high density-connectivity of clusters.

Summary

FairDen is the first **density-based group-level fair clustering** algorithm. It can even cluster data with:

- multiple sensitive attributes,
- multiple sensitive groups, and
- categorical features.

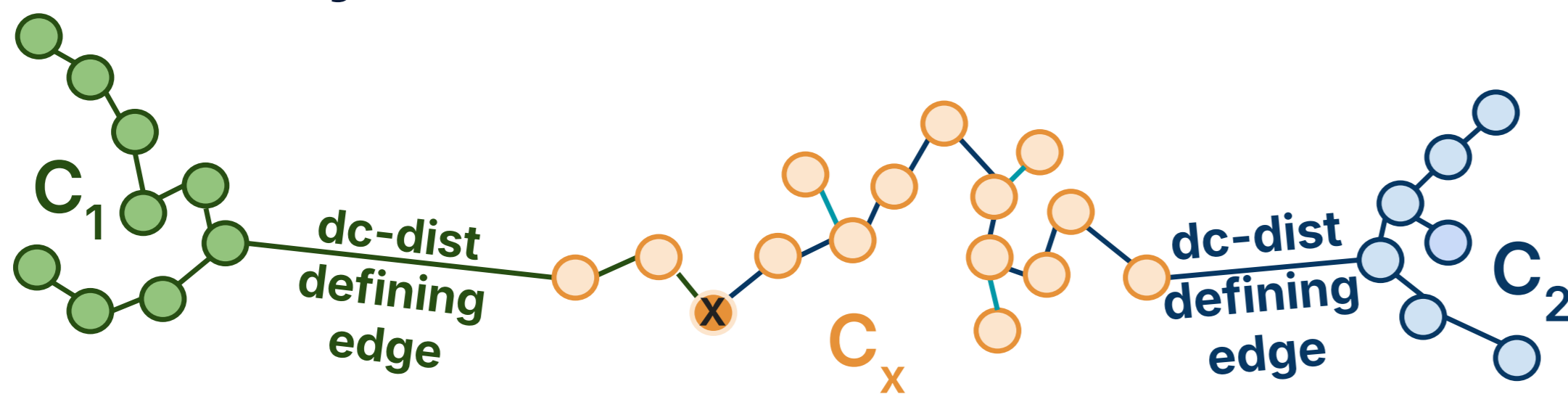
Related Work

Most other state-of-the-art group-level fair clustering methods cannot handle multiple sensitive attributes, multiple sensitive groups, or categorical values in the non-sensitive attributes.

Algorithm	Fairlet [3]	Scalable Fair Clustering [4]	Fair SC [2]	FairDen (Ours)
Density-based	×	×	×	✓
Multiple sensitive attributes	×	×	×	✓
Multiple (>2) sensitive groups	×	×	✓	✓
Categorical Features	×	×	×	✓

Method

Capture density-connected structures:



The *density-connectivity distance* [1] is based on the mutual reachability distance d_m that takes the density of points into account by using the points' core distances d_{core} :

$$d_m(x, y) = \max(d_{core}(x), d_{core}(y), d_{euct}(x, y))$$

The longest edge e along the min-max path on the graph given by d_m indicates the density-connectivity between two points x and y

$$d_{dc}(x, y) = \min_{P \in \mathcal{P}} \max_{e \in P(x, y)} |e|$$

Ensure balance of sensitive attributes:

$$f_p^{S_x} = (|S_x|/n) \cdot \mathbf{1}_n$$

a. Binary encoding for each sensitive group S_x

$$\begin{matrix} \triangle & \bigcirc & \bigcirc & \triangle \\ \triangle & (1 & 0 & 0 & 1) \\ \bigcirc & (0 & 1 & 1 & 0) \end{matrix}$$

b. Columns of fairness matrix:

$$\triangle (1 \ 0 \ 0 \ 1) - \frac{|\triangle|}{n} \cdot \mathbf{1}_n$$

Combine density-connectivity and fairness:

$$\min_{\mathcal{H} \in \mathbb{R}^{n \times k}} \text{Tr}(\mathcal{H}^T \mathcal{L} \mathcal{H}) \text{ subject to}$$

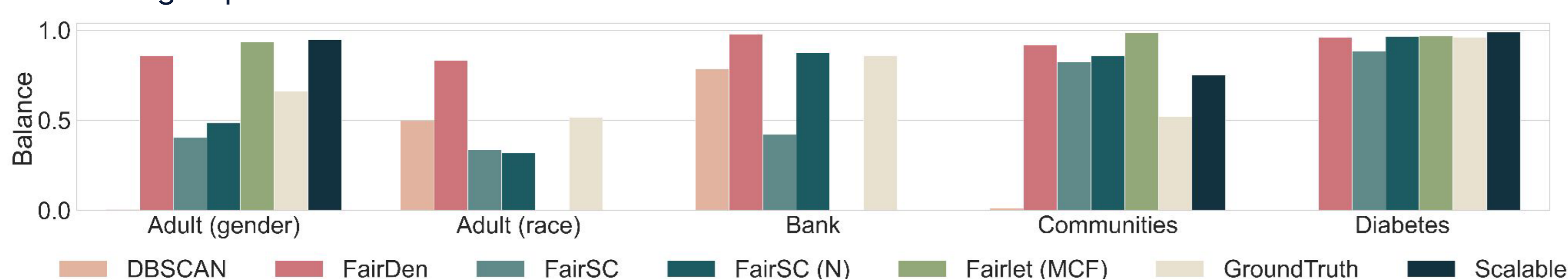
$$\mathcal{H}^T \mathcal{D} \mathcal{H} = \mathcal{I}_k \text{ and } \mathcal{F}^T \mathcal{H} = 0$$

Imposing the fairness constraint ($\mathcal{F}^T \mathcal{H}$) to a hierarchy of density-connected clusters transforms the problem into a **graph-cut problem** solvable with **spectral clustering** [2].

Results

FairDen finds more balanced clusterings with respect to sensitive attributes than other density-based methods and detects density-based clusters better than other fair methods.

Experiments regarding fairness: Balance values for all competitors on benchmark datasets. Fairlet (MCF) and Scalable Fair Clustering are not always included as they cannot handle non-binary sensitive groups.



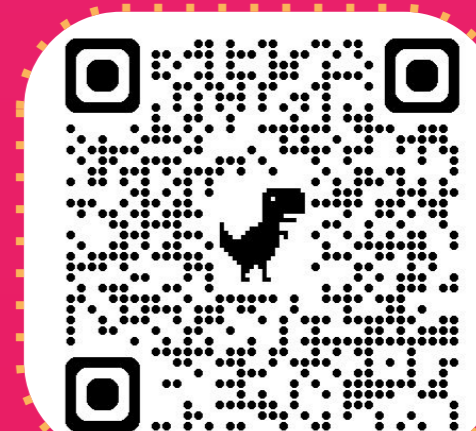
Experiments regarding clustering quality: Number of clusters k , **Balance** (assessing group-level fairness), **DCSI** (assessing density-connectivity of clusters), and **ARI** (assessing similarity to ground truth clustering) for real-world benchmark data.

k	Algorithm	Balance	DCSI	ARI
2	DBSCAN	0.01	0.97	0.00
2	FairDen	0.86	0.04	0.05
2	FairSC	0.40	0.00	0.23
2	FairSC (N)	0.49	0.00	0.27
2	Fairlet (MCF)	0.94	0.00	0.00
2	Scalable	0.66	0.00	1.00
2	GroundTruth	0.95	0.01	-0.01
2	DBSCAN	0.50	0.99	0.02
2	FairDen	0.83	0.09	0.05
2	FairSC	0.34	0.00	-0.03
2	FairSC (N)	0.32	0.00	0.16
2	Fairlet (MCF)	-	-	-
2	Scalable	-	-	-
2	GroundTruth	0.52	0.00	1.00
2	DBSCAN	0.79	0.99	0.01
2	FairDen	0.98	0.14	0.21
2	FairSC	0.42	0.00	-0.06
2	FairSC (N)	0.88	0.00	-0.04
2	Fairlet (MCF)	-	-	-
2	Scalable	-	-	-
2	GroundTruth	0.86	0.00	1.00

References

- [1] Anna Beer, Andrew Draganov, Ellen Hohma, Philipp Jahn, Christian MM Frey, and Ira Assent. "Connecting the Dots--Density-Connectivity Distance unifies DBSCAN, k-Center and Spectral Clustering." In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 80-92. 2023.
- [2] Matthäus Kleindessner, Samira Samadi, Pranjal Awasthi, and Jamie Morgenstern. "Guarantees for spectral clustering with fairness constraints." In *International conference on machine learning*, pp. 3458-3467. PMLR, 2019.
- [3] Chierichetti, F., Kumar, R., Lattanzi, S., & Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in neural information processing systems*, 30.
- [4] Backurs, A., Indyk, P., Onak, K., Schieber, B., Vakilian, A., & Wagner, T. (2019, May). Scalable fair clustering. In *International conference on machine learning* (pp. 405-413). PMLR.

Code



Paper

